Speaker Independent Acoustic-to-Articulatory Inversion

An Ji
Marquette University

SPEAKER INDEPENDENT ACOUSTIC-TO-ARTICULATORY INVERSION

by

An Ji, BSEE, MSEE

A dissertation submitted to the Department of Electrical and Computer Engineering, in Partial Fulfillment of Requirements for the Degree of Doctor of Philosophy at Marquette University

Milwaukee, Wisconsin

December 2014

ABSTRACT

SPEAKER INDEPENDENT ACOUSTIC-TO-ARTICULATORY INVERSION

Acoustic-to-articulatory inversion, the determination of articulatory parameters from acoustic signals, is a difficult but important problem for many speech processing applications, such as automatic speech recognition (ASR) and computer aided pronunciation training (CAPT). In recent years, several approaches have been successfully implemented for speaker dependent models with parallel acoustic and kinematic training data. However, in many practical applications inversion is needed for new speakers for whom no articulatory data is available. In order to address this problem, this dissertation introduces a novel speaker adaptation approach called Parallel Reference Speaker Weighting (PRSW), based on parallel acoustic and articulatory Hidden Markov Models (HMM). This approach uses a robust normalized articulatory space and palate referenced articulatory features combined with speaker-weighted adaptation to form an inversion mapping for new speakers that can accurately estimate articulatory trajectories. The proposed PRSW method is evaluated on the newly collected Marquette electromagnetic articulography – Mandarin Accented English (EMA-MAE) corpus using 20 native English speakers. Cross-speaker inversion results show that given a good selection of reference speakers with consistent acoustic and articulatory patterns, the PRSW approach gives good speaker independent inversion performance even without kinematic training data.

ACKNOWLEGMENT

This 5 year PhD journey has been the most important time during my life. I would never have been able to finish my dissertation without the guidance of my advisor and committee members, help from friends, and support from my family and husband.

I would like to express my deepest gratitude to my advisor, Dr. Michael Johnson, for his guidance, caring, patience and providing me with an excellent atmosphere for doing research. From initiating and developing to finalizing the whole research, Dr. Johnson devoted time and attention to help me all along the way. I sincerely appreciate for every great discussion which have been invaluable. I would also like to thank Dr. Jeffrey Berry for guiding me to finish the Marquette EMA-MAT dataset collection and helping me to develop my background in speech pathology and audiology.

I would like to thank all my committee members Dr. Edwin Yaz, Dr. Robert Scheidt and Dr. Richard Povinelli for their valuable reviews and comments. I would like to thank all my friends at the Speech and Signal Processing Laboratory for their suggestions and help during the five years.

Finally, I would like to thank my family and husband. They are always there cheering me up and stood by me through the good and bad times.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Problem description

Human speech is generated through the movement of a complex set of articulators, including the tongue, jaw and lips, controlled together through the speech production process. Our brain has a well-developed speech region to convert basic units (phonemes) to nerve impulses, which control muscular contractions. These contractions generate a series of articulatory movements to shape the acoustic waveform. This relationship between articulatory movements and acoustics is learned through experience, such as the process of infants imitating speech or foreign language learners learning new pronunciations. This learning process includes auditory processing, acoustic and linguistic perception and articulatory motor control.

Reversing the process to estimate articulatory movements from a speech signal, known as acoustic-to-articulatory inversion, can help us understand speech production and has application to many important speech technologies. For example, articulatory information can be integrated with acoustic features to improve the performance of automatic speech recognition system (Mitra, Nam, Espy-Wilson, Saltzman, & Goldstein, 2010; Sun & Deng, 2002). Articulatory information can be used to improve the quality of the synthesized voice in speech synthesis (Ling, Richmond, Yamagishi, & Wang, 2009) and to automate the facial animation of virtual characters in films and video-games (Hofer & Richmond, 2010). Visualizing the position of the articulators from acoustic signal would be extremely useful in speech therapy systems and in Computer Aided

Language Learning (CALL) and Computer Aided Pronunciation Training (CAPT) systems.

One motivating aspect of this work is the application of acoustic-to-articulatory inversion to CALL and CAPT systems, where a reliable inverse mapping to estimate articulatory movements would be able to more accurately analyze pronunciation errors and to assist in providing detailed corrective feedback. Current CALL and CAPT systems are limited in providing such detailed feedback, partially because this inverse problem is difficult and not yet well solved.

1.2 Motivation

The main goal of this dissertation is the creation of robust and accurate models for speaker independent acoustic-to-articulatory inversion. While there has been significant prior work in articulator-to-acoustic modeling, current methods, described more fully in Section 2.5, must be trained on simultaneous acoustic and articulatory kinematic data for each speaker. However, in many applications, it is not feasible to collect such data for each user. In these cases, an efficient acoustic-to-articulatory inversion procedure needs to be developed which is robust to the lack of kinematic training data. This is important in applications such as CALL and CAPT where models learned without kinematic data are essential.

The complexities of inter-speaker differences in both articulatory and acoustic spaces result in the need to develop more sophisticated methods for normalization of multiple speakers' articulatory measurements to represent a single generalized articulatory space, for creation of speaker dependent acoustic-articulatory models, and

subsequently for adapting these models to provide accurate acoustic-articulatory mappings for new speakers for whom there is acoustic but no kinematic data. The work described here addresses the above research problems with the goal of creating a speaker independent articulatory-acoustic inversion algorithm.

1.3 Objectives and Contributions

Current approaches for estimating articulatory parameters are speaker-dependent, requiring matched kinematic and acoustic data for the specific target speaker. Developing speaker independent methods for speech inversion is essential to furthering research in this area. The objective of this dissertation is to extend current methods for acoustic-articulatory inversion to work on new speakers with no kinematic data and limited acoustic data. Successful achievement of this objective requires advances in techniques for articulator space normalization and the application of current methods for speaker adaption to the problem of acoustic-articulatory inversion. This work has resulted in several distinct contributions:

1. **The Marquette University Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus**. The first contribution is the collection and dissemination of a multi-speaker EMA data set. This data set is one of the largest of its kind, providing simultaneous kinematic and acoustic data from 40 gender and dialect balanced speakers.

2. **A new method for articulatory space calibration**. The second contribution is a calibration approach for transformation of kinematic data into an appropriate and stable articulatory coordinate space. Results show that this calibration method

accurately and consistently transforms sensor data into an articulatory space in which sensor movements and orientations have a consistent representation. This representation enables investigation of the relationship between articulator kinematics and acoustics across speakers within a consistent articulatory space.

3. **Palate referenced articulatory features for vocal tract modeling**: The third contribution is the introduction of a set of articulator feature variables that are palate referenced and normalized with respect to the articulatory space. The selection of effective articulatory features is an important component of acoustic-to-articulator inversion and articulatory synthesis. Although it is common to use direct articulatory sensor measurements as feature variables, this approach fails to incorporate important physiological information such as palate height and shape and thus is not as representative of vocal tract cross section as desired. The features introduced here include normalized horizontal positions and normalized palatal height of two midsagittal and one lateral tongue sensor, as well as normalized lip separation and lip protrusion. The quality of the feature representation is evaluated qualitatively by comparing the variances and vowel separation in the working space and quantitatively through measurement of acoustic-to-articulator inversion error. Results indicate that the palate-referenced features have reduced variance and increased separation between vowels spaces and substantially lower inversion error than direct sensor measures.

4. **A novel speaker independent acoustic-to-articulatory inversion system works on new speakers for whom there is no kinematic data**: The fourth contribution is the Parallel Reference Speaker Weighting (PRSW) Hidden Markov Model

(HMM)-inversion system, which can adapt to new speakers without any kinematic data. By adapting in acoustic space, an adapted parallel articulatory model can be estimated to perform the inversion. Experimental results show that the PRSW approach offers good speaker independent inversion performance without kinematic training data, but requires a carefully chosen set of reference speakers with a consistent within speaker acoustic-to-articulatory mapping.

1.4 Dissertation outline

Chapter 2 starts with general background related to this work. Articulatory data, speech and articulator modeling, current articulatory normalization, speaker adaptation and previous work in the area of speech inversion have been discussed.

Chapter 3 provides a detailed description of the Marquette EMA-MAE dataset and introduces the first three contributions, including new approaches for bite-plate space calibration and palate referenced articulatory feature extraction.

Chapter 4 introduces the baseline Hidden Markov inversion model and tuning of model parameters to achieve the highest inversion accuracy. This baseline inversion system is used as an evaluation platform for the proposed articulatory features.

Chapter 5 describes the proposed PRSW method, presents an evaluation framework, and presents results of the final system on new speakers trained without kinematic data.

Chapter 6 gives conclusions and possibilities for future work

2 Background

2.1 Introduction

Humans produce audible speech by moving their articulators, particularly the tongue, lips and jaw, to modify the glottal source energy. Speech inversion aims to invert this process and determine the underlying articulatory space configuration from acoustic speech. The recovery of the articulatory movement from the acoustic signal has attracted the interest of researchers because a successful solution to this inversion problem would have many speech applications including automatic speech recognition, speech synthesis and pronunciation training. In order to solve the speech inversion problem, it is important to understand speech and articulatory data representation and basic speech modeling methods. This chapter provides a general technical background for the speech inversion research area, including articulatory data acquisition and articulatory space representation, speech acoustic modeling, speaker adaptation and previous work on acoustic-to-articulatory inversion.

2.2 Articulatory data acquisition and space representation

2.2.1 Articulatory data acquisition

There are several approaches to collecting articulatory kinematic data, including X-ray cinematography, cine MRI, ultrasound and electromagnetic articulography (EMA). Each has advantages and disadvantages related to factors such as spatial and temporal resolution, accuracy, capacity and accessibility. X-ray cinematography uses x-ray film

photography to provide accurate imaging of the articulators; however, there are concerns about radiation to the subject's head (Houde, 1967; Munhall, Vatikiotis-Bateson, & Tohkura, 1998). Magnetic Resonance Imaging (MRI) uses a magnetic field and pulses of radio wave energy to take images of structures inside the body. It can provide dynamic 3D measurement of the vocal tract but it is cumbersome and expensive (Masaki et al., 1999; Narayanan, Nayak, Lee, Sethy, & Byrd, 2004). In contrast, the ultrasound technique, which uses high-frequency sound waves to view soft tissues, is able to capture the surface of the tongue (Kaburagi & Honda, 1994; Stone, Sonies, Shawker, Weiss, & Nadel, 1983) but noise, echo artifacts and refractions may affect the results.

Electromagnetic articulography (EMA) sensing has become the most widely used articulography technique for the collection of parallel acoustic and articulatory data (Perkell & Cohen, 1992). This technique uses electromagnetic transducer coils glued to the articulators to record measurement of their position. Compared to the other techniques, EMA is low cost and relatively simple to use.

With the development of these data collection techniques, several parallel acoustic-to-articulatory datasets have become available to the public research community These include the X-ray Micro-beam Speech Production database (Westbury, 1994a), the MOCHA TIMIT database (Richmond, Hoole, & King, 2011; Wrench & William, 2000), the EUR-ACCOR multi-language articulatory database (Wrench, 1993) and the recent Edinburgh speech production facility DoubleTalk corpus (Scobbie et al., 2013). However these database are limited in the number of speakers, which makes investigation of speaker independent acoustic-to-articulatory inversion, a central component of this work, infeasible.

To address this limitation, it was necessary to collect a new multiple speaker dataset. The EMA-MAE corpus, a new bilingual multi-speaker corpus of parallel acoustic and EMA kinematic data have been collected and use it in this work to develop and test a new speaker independent acoustic-to-articulatory inversion method. A detailed description of this corpus will be given in Chapter 3.

2.2.2 Articulatory space representation

Representation of articulatory motion is very important in acoustic-to-articulatory inversion. Currently, most approaches have suggested that linguistically based features which relate directly to the human articulatory process, such as tongue position, lip rounding, place of articulation, and manner of articulation, can be beneficial in capturing speech characteristics (Kirchhoff, 1999; Metze & Waibel, 2002; Tang, Seneff, & Zue, 2003). These articulatory features are abstract descriptions of vocal tract properties and articulator motion during speech production; therefore they can complement or even replace acoustic-based features in speech processing. Recently, there has been renewed interest in applying articulator information as alternative and or supplementary features for speech processing tasks (Erler & Deng, 1993; Frankel & King, 2001; Leung & Siu, 2004). While there is general agreement on the articulator properties of base phonemic units, there are many ways to represent or encode these properties such that they can be extracted and modeled with no standard representation. There are a number of different articulatory models that have been proposed (Birkholz, Jackel, & Kroger, 2006; Coker, 1976; Mermelstein, 1973). The Maeda model (Maeda, 1990) shown in Figure 2.1 is a common model which represents the articulatory space and motion with seven key

parameters that relate to the cross-sectional area of the vocal tract, originally constructed by applying a factor analysis method on vocal tract contour data.

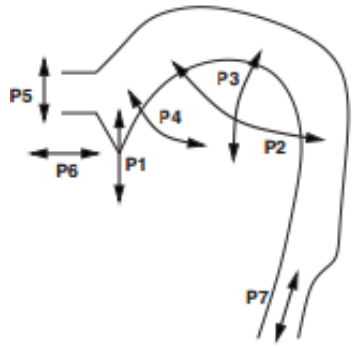

Figure 2.1: Maeda's articulatory model: P1 jaw height, P2 tongue dorsum length, P3 tongue dorsum shape, P4 tongue apex position, P5 lip separation, P6 lip protrusion, P7 larynx height

These established models typically represent a two-dimensional midsagittal vocal tract, and do not include modeling of more complex features such as lip rounding or tongue curvature. Three-dimensional articulatory modeling, specifically including 3D tongue reconstruction, lip position and facial shape, has seen noteworthy advances (Badin et al., 2002; Birkholz et al., 2006; Dang & Honda, 2004; Story, 2005). Detailed 3D knowledge of the vocal tract shape is important for more realistic speech production studies. However, these efforts all require the combination of multiple medical imaging techniques to provide complementary spatial and temporal resolution of variation in the vocal tract. Moreover, the high-dimensional nature of three-dimensional articulatory models substantially complicates speaker normalization and acoustic-to-articulatory

inversion. Consequently, three-dimensional vocal tract modeling is largely constrained to speaker-specific models and has not yet become accessible for multi-speaker research.

In this work a Cartesian coordinate system is used for the articulatory space, referenced to each individual subject's physiology such that the midsaggital plane and maxillary occlusal plane form the axes of the articulatory space, as described in more detail in Section 3.3. Within this articulatory space, each subject has an unique dynamic range of motion, creating what is referred to as their "articulatory working space", "working space", or "vowel space", since much of this dynamic range, especially of the tongue, is a function of vowel-related motion.

2.2.3 Articulatory space normalization

In order to make meaningful comparisons across speakers in the articulatory space or to develop robust speaker independent acoustic-to-articulatory inversion systems, normalization in the articulatory space across speakers is necessary. Both articulatory and acoustic structure vary substantially across speakers due to physiological differences as well as learned language and dialectal pronunciation differences. Understanding the source of speaker variability is important when designing a procedure that recovers articulatory movement from speech acoustics.

Hashi proposed a geometric-based normalization method for articulatory parameters. In his paper, the palatal height is used as a systematic source of variation and the articulatory data is scaled to a common range. Specifically the tongue and lip positions are expressed relative to the normalized palate. Sadao has also implemented palate normalization, performed by rotating the palate positions for the position of upper

incisor. Rotation angle is determined by minimizing the error of palate positions among different speakers (Hiroya & Mochida, 2005). McGowan and Cushing proposed vocal tract normalization for articulatory inversion using analysis-by-synthesis. In their work the normalization is implemented by adjusting the articulatory model in order to make the acoustic signal and articulatory model match as closely as possible over pairs of corresponding human and model midsagittal shapes (McGowan & Cushing, 1999). Felps and Osuna (Felps & Osuna, 2010) describe and compare two articulatory normalization methods across speakers, the classical and extended Procrustes transformation, which allows for global translation, rotation and scaling of articulator positions. Results indicate that the extended Procrustes with an analysis-by-synthesis loop can find an optimized articulatory normalization space with consistent acoustic similarity.

The ideal normalization method is largely dependent on the corpus and target application, so there is no consensus on which is the best normalization method. For example, Bechman et al. straighten the vocal tract wall to transform the coordinates for MRI data (Beckman & Jung, 1995). Hashi et al normalize the vowel posture in the X-ray Microbeam database (Hashi, M. Westbury, J. R. & Honda, 1998). Wei et al use thin-plate splines to reduce the morphological differences of vocal tracts among different subjects with EMA data (Wei, 2008). All of these normalization methods work for a specific dataset but are not necessarily broadly applicable to all kinematic measures.

In this dissertation, a geometric based articulatory space calibration and normalization is used for the Marquette MAE-EMA corpus and the speaker independent acoustic-to-articulatory task, as described in detail in Section 3.3. From this articulatory space, a set of articulatory feature variables are computed, which incorporate range of

motion and palate information to further normalize the final representation of articulatory motion across speakers, as described in Section 3.4.

2.3 Speech acoustic modeling

2.3.1 Acoustic features

The previous section presented the representation of articulatory features varies across different tasks. In contrast, the typical representation of speech is relatively consistent. Normally most inversion systems use standard Cepstrum analysis (Davis & Mermelstein, 1980) to generate a set of features, called Mel Frequency Cepstral Coefficients (MFCCs), which are a robust representation of vocal tract configuration information regardless the source of excitation. This feature is also the most commonly used feature in automatic speech recognition systems. Some inversion systems use Linear Predictive Coding (Lawrence & Schafer, 1978) coefficients and Perceptual Linear Prediction (Hermansky, 1990), but these representations have been generally replaced by MFCCs. This work uses MFCCs and MFCC dynamics (velocity and acceleration) as acoustic features.

2.3.2 Statistical acoustic modeling

Acoustic modeling of speech is the process of capturing the relationship between sound units and acoustic feature vectors. The acoustic input consists of a sequence of feature vector observations O. Each index represents a discrete time interval, and successive o_i indicate temporally consecutive frames of the input:

$$O = [o_1, o_2, o_3, \dots, o_T]. \tag{2.1}$$

Similarly, we can represent the sequence of sound units as

$$W = [w_1, w_2, w_3, \dots, w_n]. \tag{2.2}$$

In the context of automatic speech recognition, the goal is to find the most likely sound unit sequence given the acoustic input O:

$$\widehat{W} = argmax(P(W|O)). \tag{2.3}$$

By using Bayes' rule we can break the above equation down as

$$\widehat{W} = argmax(\frac{P(O|W)P(W)}{P(O)}), \tag{2.4}$$

Here $P(W)$ is the prior probability of the unit sequence, computed from a language model. $P(O|W)$ is the observation likelihood from the acoustic model. $P(O)$, the probability of the acoustic observation sequence, which for maximum likelihood estimation of W, is not needed:

$$\widehat{W} = argmax\left(\frac{P(O|W)P(W)}{P(O)}\right) = argmax(P(O|W)P(W)). \tag{2.5}$$

Since the true alignment between W and O is unknown even in labeled training data, the underlying state sequence is 'hidden', and an appropriate model choice is a discrete state statistical state machine, such as Hidden Markov Models (HMM). An HMM consists of two stochastic processes, a hidden Markov chain and an observable process. Figure 2.2 shows a left-to-right 6-state HMM structure for acoustic modeling.

The parameters needed to define the HMM are:

- **States:** a set of states ($S_1 - S_n$)

- **Transition probabilities**: a set of probabilities $A = [a_{11}a_{12} \ldots a_{n1} \ldots a_{nn}]$. Each a_{ij} represents the probability of transitioning from state i to state j.
- **Observation likelihoods:** a set of observation likelihoods $B = b_i(o_t)$, each represents the probability of an observation o_t being generated from a state i
- **Initial distribution:** an initial probability distribution over the states, such that π_i is the probability that the HMM will start in state i.

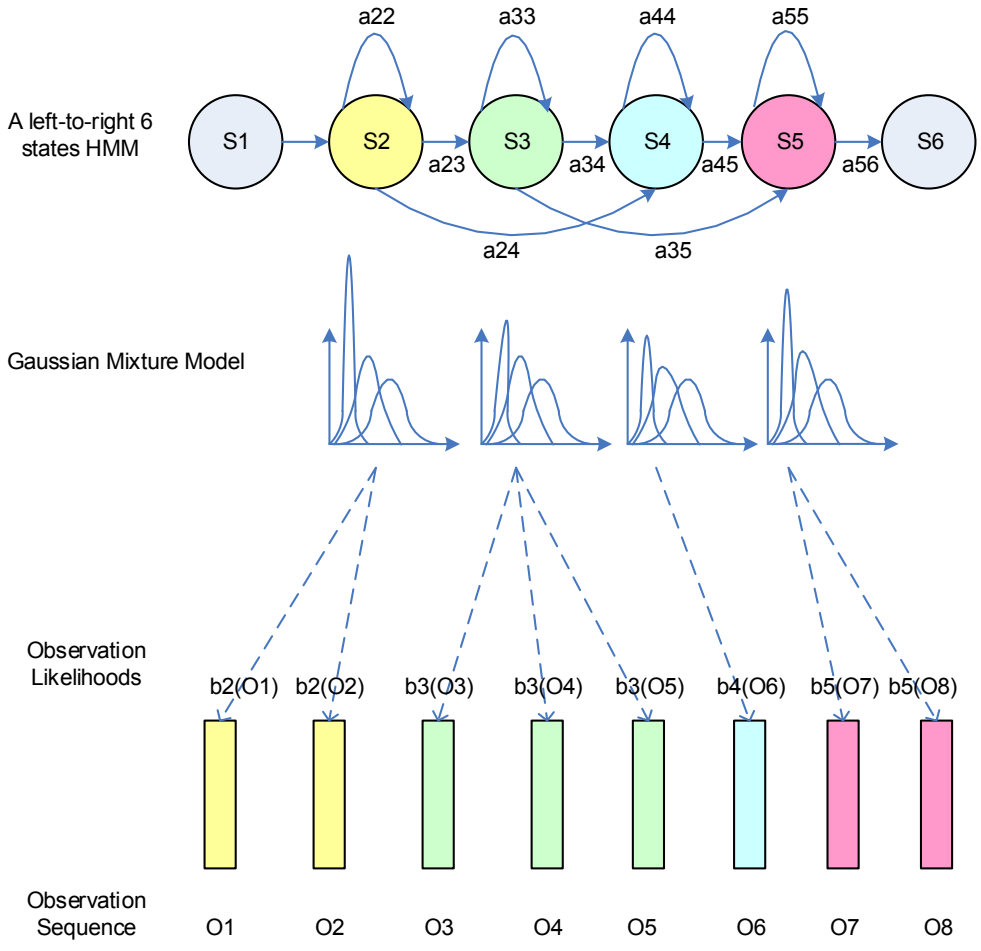

Figure 2.2 Left-to-right 6 state HMM structure

In the above HMM example, there are two special states called non-emitting states used as the start (S1) and end state (S6), which allow for connecting multiple HMMs together in a longer sequence. At each time interval t within i state, an observation feature vector O is generated by the probability density function $b_i(o_t)$. All states generate observations except the two non-emitting states. The observation distribution $b_i(o_t)$ is typically represented by Gaussian mixture models (GMMs):

$$b_i(o_t) = \sum_{m=1}^{M_i} c_{im} N(o_t;\ \mu_{im}, \sum_{im}), \tag{2.6}$$

where M_i is the number of mixture components for state i, and c_{im} is the weight of component m of state i. $N(o_t;\ \mu_{im}, \sum_{im})$ is the m th mixture normal density function of state i:

$$N(o_t;\ \mu_{im}, \sum_{im}) \propto |\sum_{im}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(o_t - \mu_{im})^T \sum_{im}^{-1}(o_t - \mu_{im})\right) \tag{2.7}$$

HMMs have been the dominant acoustic model for speech recognition for nearly 30 years (Jelinek, 1999; Rabiner, 1989; Rabiner & Juang, 1993). The basic inversion framework in this dissertation is based on HMM acoustic modeling.

There have been several other models proposed for acoustic modeling in recent studies, such as conditional random fields, artificial neural networks, hidden/linear dynamic models and others (Bahl & Jelinek, 1975; Jelinek, 1969; Jelinek, Bahl, & Mercer, 1975; Jelinek, 1976). These models have advantages for specific applications, but HMMs remain the most widely used approach.

2.4 Speaker adaptation

The main goal of this dissertation is find a robust speaker independent inversion mapping to estimate a new speaker's articulatory trajectory without any kinematic training data. To do this, existing model based speaker adaption methods used for speech recognition can be utilized. The idea of adaptation is to create a new acoustic model for the target speaker from existing trained reference speaker models, with a minimal amount of training data for the new speaker, called the adaptation data. Normal adaptation algorithms include Bayesian-based maximum *a posteriori* (MAP) (Gauvain & Lee, 1994), the transformation-based maximum likelihood linear regression (MLLR) (Leggetter & Woodland, 1995), Reference Speaker Weighting (RSW) (Hazon & Glass, 1997; Hazon, 2000) and Eigenvoice (Kuhn, 1998; Kuhn, Junqua, Nguyen, & Niedzielski, 2000).

By using acoustic adaption techniques, we intend to identify differences in acoustic patterns and create adapted acoustic and kinematic models in parallel, and form a new inversion mapping that can estimate articulatory trajectory on new speakers with no kinematic data. The MAP and MLLR methods are not suitable for adapting articulatory models directly from acoustic models because there is no kinematic data available for the target speaker to perform articulatory adaptation. In the context of acoustic-to-articulatory inversion, the idea behind RSW is more appropriate because this assumes that the model parameters of a new speaker can be constructed from a weighted combination of a set of individual reference speakers' models. This combination can be extended to the articulatory space to develop a speaker independent inverse mapping.

Since the proposed new method is based on the RSW concept, we will elaborate the technical details of RSW in more detail in this section.

2.4.1 Reference Speaker Weighting (RSW)

Rapid speaker adaptation approach implements adaptation with very small amounts of adaption data, typically 5-10 seconds of speech (Kubala, Schwartz, & Barry, 1989). Reference speaker weighting is based on model combination and works effectively even when the amount of adaptation data is quite small. RSW requires speaker-dependent models as a starting point for estimating the parameters of a new speaker.

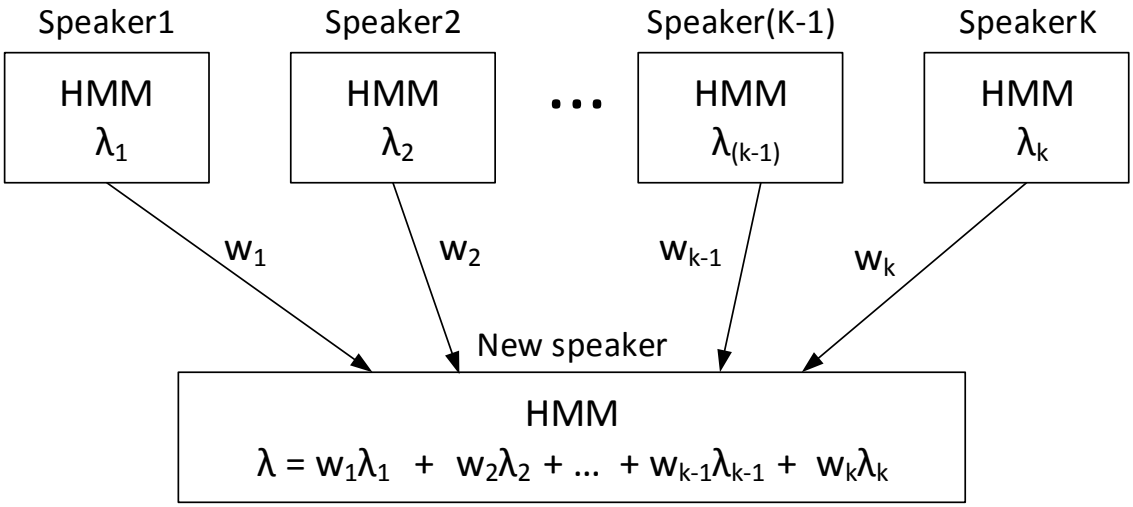

Figure 2.3 Reference speaker weighting

The basic idea of this method is shown in Figure 2.3 and 2.4. A new speaker's model can be estimated from a weighted combination of reference speakers. Each reference speaker is represented by a supervector, which is constructed by concatenating the mean vectors of all acoustic model parameters.

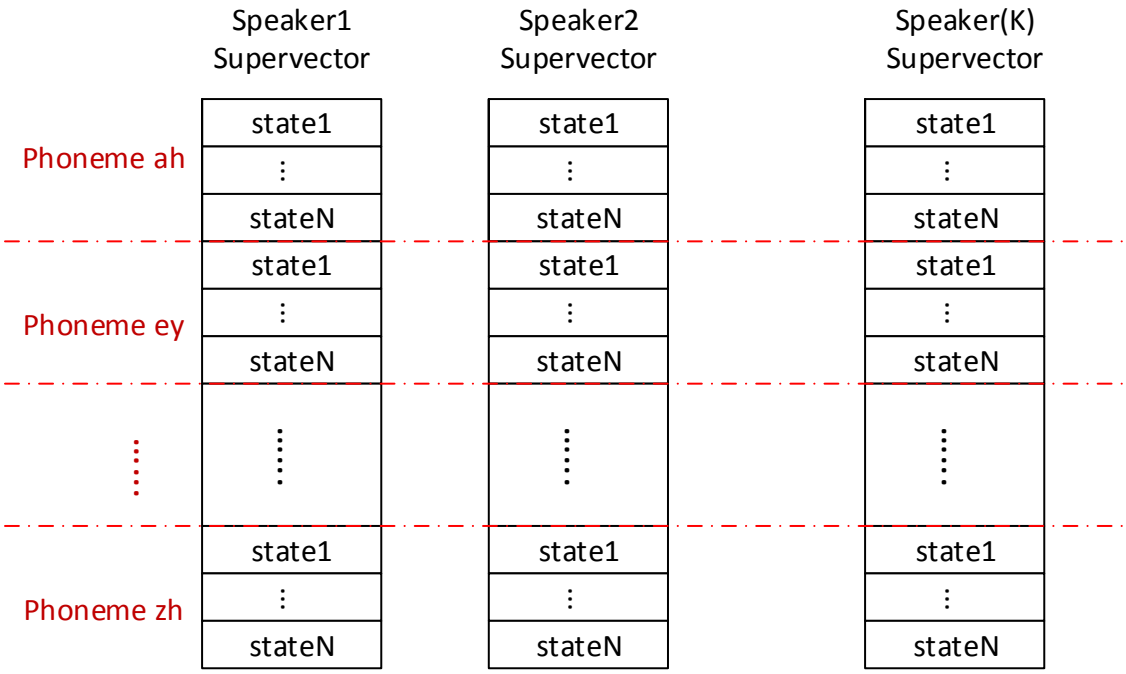

Figure 2.4 Supervector representation of reference speakers

RSW estimates the model of a new speaker from the span of the reference speaker models. Figure 2.5 shows the implementation procedures of RSW in the acoustic space. In the offline steps, speaker dependent models are trained using HMMs. Supervectors are used to represent the HMM model parameters. Once the reference speakers' models are constructed, the online steps estimate weights from new speaker's adaptation data by

using the expectation maximization algorithm to determine maximum likelihood weight estimates. The new speaker's model can then be constructed from a linear combination of reference speakers' model using these weights.

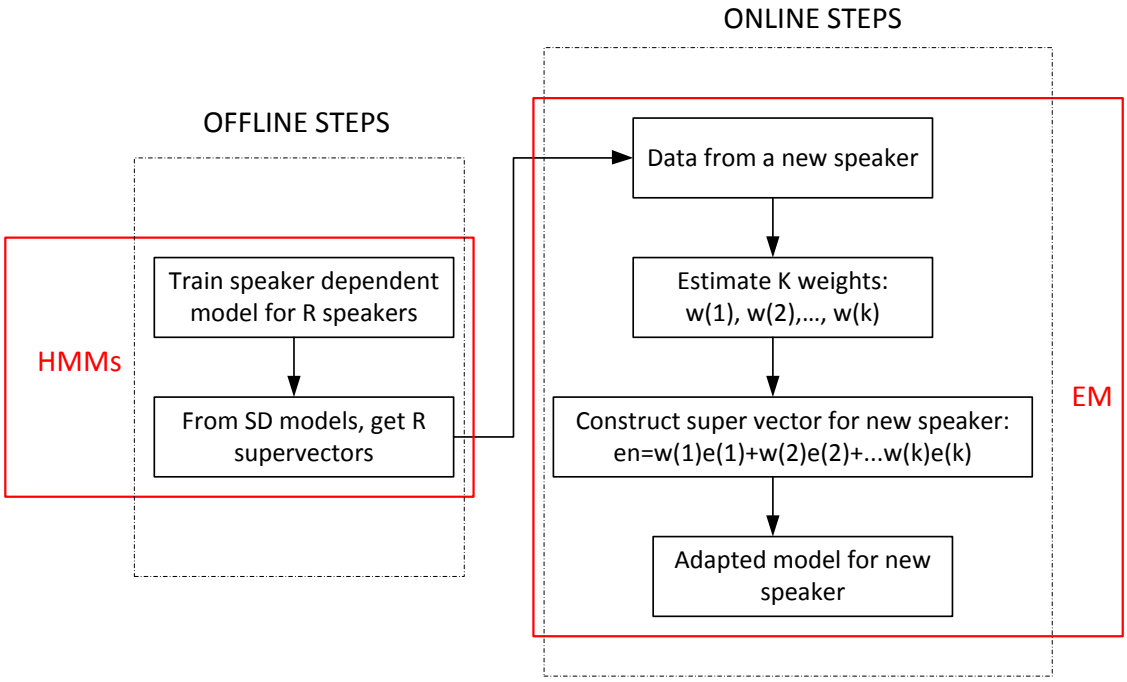

Figure 2.5 RSW implementation diagram

The weights are estimated by comparing the acoustic signal from a new speaker S against a set of K reference speakers. Let $Y = \{y_1, y_2, \dots, y_K\}$ be the set of reference speaker supervectors, defined as the concatenation of the Gaussian means from all state models in sequence. Then the RSW estimate of the new speaker's supervector is

$$s \approx s^{rsw} = \sum_{k=1}^{K} w_k y_k = YW \qquad (2.8)$$

and the mean vector of the rth Gaussian is

$$\mu_r^{(rsw)} = \sum_{k=1}^{K} w_k y_{mr} = Y_r W, \tag{2.9}$$

where $W = [w_1, w_2, \dots, w_K]'$ is the weight vector and r is the number of Gaussian mixtures

Given the adaptation data $O = \{o_t, t = 1, \dots, T\}$, the Maximum Likelihood estimate of w can be found by maximizing the following $Q(w)$ function:

$$Q(w) = -\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t(r) \left(o_t - s_r^{(rsw)}(w)\right)' C_r^{-1} (o_t - s_r^{(rsw)}(w)) \tag{2.10}$$

where $\gamma_t(r)$ is the posterior probability of observing o_t in the r_{th} Gaussian, and $\mathrm{C_r}$ is the covariance matric of the rth Gaussian. The optimal weight vector may be found by setting

$$\frac{\partial \mathrm{Q}}{\partial \mathrm{w}} = 2 \sum_{\mathrm{r}=1}^{\mathrm{R}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \gamma_{\mathrm{t}}(\mathrm{t}) \mathrm{Y}_{\mathrm{r}}' \mathrm{C}_{\mathrm{r}}^{-1} (\mathrm{o}_{\mathrm{t}} - \mathrm{Y}_{\mathrm{r}} \mathrm{w}) = 0 \, . \tag{2.11}$$

Thus, the weights w may be obtained by solving a system of K linear equations,

$$\mathrm{w} = [\sum_{\mathrm{r}=1}^{\mathrm{R}} (\sum_{\mathrm{t}=1}^{\mathrm{T}} \gamma_{\mathrm{t}}(\mathrm{r})) \mathrm{Y}_{\mathrm{r}}' \mathrm{C}_{\mathrm{r}}^{-1} \mathrm{Y}_{\mathrm{r}}]^{-1} [\sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{Y}_{\mathrm{r}}' \mathrm{C}_{\mathrm{r}}^{-1} (\sum_{\mathrm{t}=1}^{\mathrm{T}} \gamma_{\mathrm{t}}(\mathrm{r}) \mathrm{o}_{\mathrm{t}})] \tag{2.12}$$

RSW uses the model parameters of selected speakers to create a composite model for new unseen speakers. Another fast speaker adaptation method which is very similar to RSW is Eigenvoice. Eigenvoice uses principal component analysis to find a set of orthogonal basis vectors to create reference vectors. Both of these methods require the model of a new speaker to lie on the span of some reference vectors. The only difference is in the ways that the reference vectors are computed. In our acoustic-to-articulatory

inversion application, RSW is chosen because we have one-to-one matched acoustic and articulatory models for individual speakers, which allows us to use the information from the acoustic space to adapt the model in articulatory space.

2.5 Previous work in speech inversion

In the previous sections we have reviewed articulator and acoustic modeling, and reference speaker weighting adaptation. In this section, previous work on acoustic-to-articulatory inversion will be discussed.

Conversion between acoustic and articulatory representations of the vocal tract is not an easy task. The transformation of acoustic data into an articulatory feature representation is not yet solved (Laprie, 1998), although several methods have been proposed. One of the reasons for the difficulty is the "one to many" problem: a given articulator state has only one acoustic realization but this acoustic signal can be the outcome of more than one articulator state. The non-uniqueness of the mapping between acoustics and articulation has been observed by many researchers. Lindblom (Lindblom, Lubker, & Gay, 1977) with his colleagues found that subjects were able to generate formants within the ranges of variation of normal vowels in spite of physiologically unnatural jaw openings from bite-block experiments. The bite-block experiments asked subjects to produce Swedish vowels with constrained and unconstrained mandible in Atal 's (Atal, Chang, Mathews, & Tukey, 1978) study of relationships between the shape of the vocal tract and its acoustic realization. They observed that the shape of the vocal tract can be changed without changing the formant frequencies. Different vocal tract shapes can generate near identical values for the first three formant frequencies. For example,

the English vowel /i/ can be produced with several positions while keeping its formant characteristics. From the theoretical side, analyzing Webster's horn equation, a second-order linear differential equation used to derive transfer function of a tube under some boundary conditions, the area functions $A\ (x)$ and $1/A(L-x)$ (where L is the length of the vocal tract) produce the same acoustic signal (Qin & Carreira-Perpinan, 2007). All of these observations and findings support the non-uniqueness nature of acoustic-to-articulatory inversion.

Although this non-uniqueness is a legitimate concern, it is typically observed within a relatively small range (Qin & Carreira-Perpinan, 2007), as discussed in more detail in section 4.2. Speech inversion has the potential to benefit existing speech recognition systems, especially in cases with noisy, spontaneous, pathological or nonnative speech. In addition to automatic speech recognition, other possible applications include speech synthesis and Computer Aided Language Learning systems.

2.5.1 Codebook method

The articulatory codebook method estimates articulatory parameters by looking up pairs of segmental acoustic and articulatory features from parallel recoded articulatory-acoustic data. Hogden (Hogden et al., 1996) divided acoustic vectors into 256 codes through vector quantization by finding the shortest Euclidean distance between the acoustic vectors and articulatory vectors. Kaburagi and Honda (Kaburagi & Honda, 1998) used the codebook method to synthesize the speech spectrum. In this method each articulatory and acoustic data pair stored nine positions and the values of the line spectral pair (LSP) parameters throughout the utterance. Using Vector Quantized codebooks is a

discrete approach and cannot give a high resolution approximation without significantly increasing the size of data. Since more sophisticated statistical models have been developed, this method has largely been replaced.

2.5.2 Neural network method

Richmond (Richmond, 2002) proposed a successful mapping of the speech signal onto EMA data by using Neural Networks, including Multiplayer Perceptrons and Mixture Density Networks. He obtained good inversion results with 1.40mm RMS error for two MOCHA-TIMIT EMA speakers. The neural network method has shown to be an accurate model for inverse mapping if given enough data. An inversion system based on neural networks is straightforward to implement, but the choice of network structure, for example number of hidden layers and nodes per hidden layer requires significant tuning. In addition, phonetic or other temporal constraints cannot be easily incorporated in this approach.

2.5.3 Kalman filter

King and Wrench presented a dynamical system model using Kalman filter trained on EMA data (King & Wrench, 1999). They concluded that the underlying physical mechanism of speech production is sufficiently linear as not to require non-linear dynamical models; however, the acoustic observations do not have a linear relationship to the articulator parameters. Dusan and Deng (Dusan & Deng, 2000) employed an extended Kalman filter trained on paired acoustic-articulatory data. Different phonological models were built by applying an extended Kalman filter on each segment of speech repetitively. Articulatory trajectories were estimated by applying the

extended Kalman smoother using the parameters of the phonological models. The reported average RMS error between estimated and actual articulatory trajectories is about 2mm.

2.5.4 Gaussian mixture model

Mixture models have also been used. Modeling the joint distribution of acoustic and articulatory features with a Gaussian Mixture Model is proposed by Toda (Toda, Black, & Tokuda, 2004). The mapping function from an acoustic feature vector x_t to an articulatory feature vector y_t in time segment t is defined as

$$\hat{y}_t = \sum_{i=1}^{M} p(m_i|x_i)p(y_t|x_t, m_i), \quad (2.13)$$

where M is the total number of mixture components, $p(m_i|x_i)$ is the component weight conditioned on x_i, and $p(y_t|x_t, m_i)$ is a conditional Gaussian distribution with full covariance matrices. The set of GMMs were trained using Maximum Likelihood Estimation on the joint probability $p(x, y)$ using parallel acoustic-articulatory data. In order to get good inversion accuracy, 128 Gaussian mixture components were used in their experiments. The best performance was found when a mixture of 32 components was used.

2.5.5 Hidden Markov model inversion

Hiroya and Honda (Hiroya & Honda, 2004) recently developed a mapping algorithm using a hidden Markov model. In this approach, each phoneme is modeled by a context-dependent HMM and the optimal maximum *a posteriori* sequence of articulatory parameter estimation is computed through Viterbi alignment. HMMs of articulatory

parameters were built for each phoneme and the mapping from the articulatory to acoustic domain was approximated in a piece-wise linear form with parameters trained from the parallel acoustic-articulatory data. In the inversion stage, an HMM state sequence was derived from the speech signal via Viterbi decoding, and then articulatory feature values were estimated from the linear mapping and a smoothed output trajectory was generated. This model approximates the mapping between acoustic and articulatory domain as a linear function, which is not able to sufficiently capture the highly complicated non-linear relationship between articulatory and acoustic domains.

Rather than combing acoustic and articulatory within a joint model, Zhang proposed an inversion method using two parallel HMM models (Zhang & Renals, 2008). In this approach, acoustic and articulatory HMMs are connected through a highly abstracted phoneme level representation. Instead of seeking a direct mapping, the articulatory domain can be mapped to acoustic domain through state sequence alignment under HMM framework. In Zhang's paper, the reported RMS error is 1.705mm for speaker independent inversion. This is competitive with the lowest published errors, specifically Richmond's multiple layer perceptron method discussed above.

This approach based on parallel HMMs is well suited for implementing adaptation algorithms in a parallel fashion, allowing us to adapt articulatory models without kinematic data. In this dissertation, the HMM based inversion framework will be used and extended to work in a speaker independent manner.

2.6 Summary

This chapter has reviewed the technical background needed to develop speaker independent acoustic-to-articulatory inversion methods, including acoustic and articulatory data acquisition, modeling, speaker adaptation methods and existing inversion approaches. In this dissertation, a new acoustic-to-articulatory inversion approach is proposed based on a parallel HMM method. This approach is a HMM based framework which is suitable for developing speaker independent inversion and implementing adaptation algorithms. The remainder of this dissertation will focus on data collection, articulatory feature extraction, and implementation and evaluation of the proposed speaker independent inversion system.

3 Marquette EMA-MAE corpus and articulatory feature extraction

3.1 Introduction

This chapter describes the EMA-MAE dataset, a new multi-speaker acoustic and EMA articulatory dataset which has been collected to investigate acoustic-articulator modeling and speaker independent acoustic-to-articulatory inversion. All of the inversion experiments in chapter 4 and 5 are based on this dataset. In addition to a detailed description of this corpus, methods for articulatory data preprocessing and articulatory feature extraction will also be discussed.

The collection of this corpus has been supported by the National Science Foundation under NSF IIS-1320892.

3.2 Marquette EMA-MAE corpus

There is a significant need for more comprehensive electromagnetic articulography (EMA) datasets that can provide matched acoustic and articulatory kinematic data with good spatial and temporal resolution. To meet this need, the Marquette University Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus has been collected to provide kinematic and acoustic data from 40 gender and dialect balanced native English speakers and Mandarin accented English speakers.

3.2.1 Data collection system

A Northern Digital NDI Wave Speech Research System has been used to collect the articulatory kinematic data. The Wave system is an EMA system specifically designed for tracking articulatory movements and articulatory kinematics. It provides two kinds of sensors, 5 DOF and 6 DOF. The 5 DOF sensors allow tracking of x, y, and z spatial coordinates, as well as angular coordinates characterizing rotation about the transverse axis (pitch) and anterior–posterior axis (roll) of the sensor. The 6 DOF sensors have the added capacity for tracking angular coordinates characterizing rotation about the inferior–superior axis of the sensor (yaw). The system samples kinematic data at 400Hz, and acoustic data at 22.05 KHz. In the most commonly used configuration as well as one in our set up, a single 6 DOF sensor is used as a reference sensor with all other sensors being 5 DOF and all position and orientation data provided relative to the primary reference sensor.

3.2.2 Subjects

The EMA-MAE corpus has 40 subjects, including two primary subject groups designated L1 and L2. The L1 group consists of 10 male and 10 female native speakers of English, with an upper Midwest American English dialect background. The L2 group consists of 10 male and 10 female native speakers of Mandarin Chinese who speak English as a second language. Within the L2 group is a further dialectal division into subjects with a northern Beijing-region dialect background, and subjects with a southern Shanghai-region dialect background, with 5 male and 5 female speakers from each of these subgroups.

Subjects are between the ages of 18-40 with no history of speech, language, or hearing pathology, no history of orofacial surgery (other than typical dental extractions), and no history of use of anticonvulsant, antipsychotic, or anti-anxiety medications (as these factors may affect motor performance).

3.2.3 Speech tasks

The corpus includes approximately 45 minutes of synchronized acoustic and kinematic data for each speaker. In order to obtain necessary and sufficient data to characterize both segmental and supra-segmental variability pertinent to the characterization of English spoken by native-Mandarin talkers, as well as to complement existing databases, both word, sentence and paragraph level speech samples have been used. The word section covers the phonetic space of English vowels, using a 383 word list developed by Rogers (Rogers, 1997) to highlight primary phonemic contexts that influence intelligibility for native-Mandarin speakers of American-English. Subjects read 330 text-prompted words in single-word citation form. Words were blocked into approximately 25 words per record, to allow monitoring of sensor adhesion and give participants regular rest and adjustment periods. The TIMIT database sentences (Garofolo et al., 1993; Zue, Seneff, & Glass, 1990) and Harvard Intelligibility Sentences (IEEE subcommittee on subjective measurements IEEE recommended practices for speech quality measurements.1969) forms the basis for the sentence level speech samples. In addition, 9 contrastive stress sentences are chosen for emphasizing the use of contrastive stress in differentiating semantic form. Six paragraphs of various lengths are also included for emphasizing different aspects of speech including general intelligibility,

breath group utilization, accented-English intelligibility, speaking rate and segmental timing.

3.2.4 Data collection framework set up

The EMA-MAE corpus includes synchronous acoustic and three-dimensional kinematic articulator data. Data were collected in an acoustic booth with participants seated in a custom plastic chair designed to allow subjects to maintain a comfortable speaking posture within the electromagnetic field. Acoustic records were obtained using a cardioid pattern directional condenser microphone positioned approximately 1 meter from participants.

As shown in Figure 3.1, articulatory sensors included the jaw (MI) (lower front incisor), lower lip (LL), upper lip (UL), tongue body (TD), and tongue tip (TT), all placed in the midsagittal plane. In addition, there were two lateral sensors, one (LC) at the left corner of the mouth to help indicate lip rounding and one (LT) in the left central midpoint of the tongue body to help indicate lateral tongue curvature.

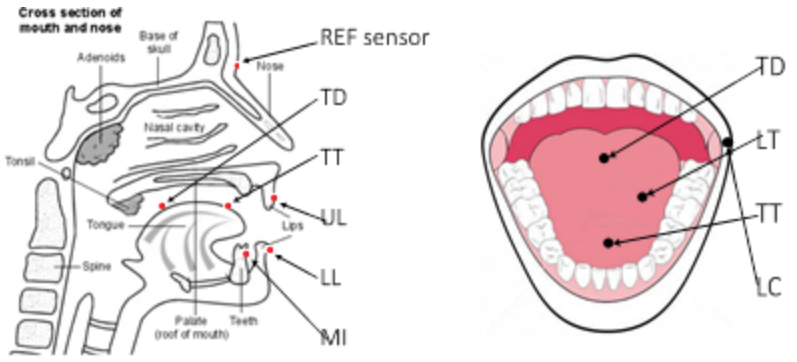

Figure 3.1 Sensor placement

A reference sensor (REF sensor) was located near the bridge of the nose using a pair of plastic glasses. The reference sensor was a 6 DOF sensor, providing three dimensional position as well as three-dimensional orientation data. All other sensors in the system were 5 DOF sensors, since these are significantly smaller and have less interference with natural subject articulation. 5 DOF sensors provide three dimensional position information but only two dimensional orientation data. This identifies the orientation, i.e., pitch and roll, of the sensor plane (which physically is a small wound toroidal coil) but no information about yaw of this plane. Position data are given in millimeters. Orientation data are given in quaternion rotation format, indicating rotation axis and angle relative to a base orientation.

Each subject underwent an initial calibration process in which softened dental wax was formed into a bite plate around a tongue depressor and a dental impression taken. Biteplate sensors are placed at the front incisor (OS) and at the mid-point of the back molars (MS) to indicate the midsaggital and maxillary occlusal planes relative to the

reference sensor, which is used to form a consistent articulatory working space. Biteplate configuration is pictured in Figure 3.2.

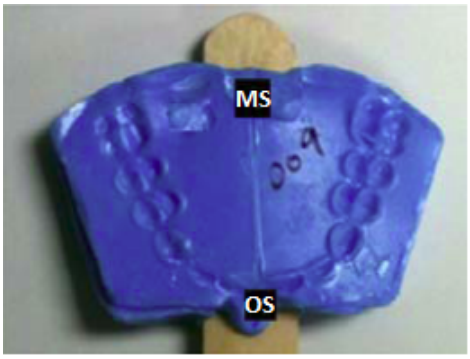

Figure 3.2 Biteplate with OS and MS sensors position

Subjects also underwent a palatal measurement in which the experimenter used a sensor-tipped palate wand to collect palatal reference data which includes a trace of the mid-sagittal palate line, a series of transverse traces across the palate, and both inner perimeter and outer perimeter dental traces at the gum line. As described in the next section, this palate information can be used to dertermine vocal tract configuration relative to tongue sensors. In addition to the biteplate and palatal measurement processes, subjects were given an acclimation period and opportunity to read some practice materials once sensors had been attached.

3.2.5 Annotation and transcription

For all subjects, a phoneme-level (broad) transcription is provided. Transcription was completed by trained graduate students in Marquette's Speech Pathology and

Audiology program using American English (IPA subset) phonemes. All transcriptions were completed by listeners with a common upper Midwest American English dialect. Multiple listener transcriptions are included for L2 subjects, to use for estimating perceived phoneme variability and perceived intelligibility. For the connected speech data, timestamps of clear pause locations (breath group and/or sentence boundaries) are included so that the paragraph-level utterances and transcriptions can be easily subdivided into sentence level data if desired.

3.3 Articulatory space calibration

3.3.1 Internal head-correction

Raw data from the EMA system are in a global coordinate space relative to the system's electromagnetic field. There is significant data processing required to compensate for subject movement and physiology to provide data in an appropriate articulatory working space.

Data is produced by the system either globally, relative to the Cartesian coordinate space established by the fixed electromagnetic field, or locally relative to the reference sensor, such that head motion is automatically removed from the data, called "head-correction". Transformation of the global coordinate data into the local coordinate space relative to the fixed reference sensor is handled in real-time by the NDI Wave software. As described in Section 3.2.4, a reference sensor mounted on a pair of plastic glasses is used with all subjects to determine and compensate for head movements.

Position data are adjusted by a direct linear translation, and orientation data are adjusted through a quaternion rotation relative to the reference sensor's orientation.

In this initial head-corrected space, the origin is at the reference sensor and the Cartesian coordinate system is relative to the orientation of the reference sensor, typically carefully placed so that the X axis represents anterior-posterior motion, the Y axis represents superior-inferior motion, and the Z axis represents lateral motion. Thus the XY plane is approximately the subject's mid-sagittal plane and the XZ plane is roughly parallel to the subject's transverse plane, but these are not exact. In order to more precisely orient the working space for each subject a bite-plate correction is implemented, as described in the next section.

To establish some measures of head correction and biteplate calibration variance, about mid-way during the data collection process an additional calibration step was added in which subjects were asked to nod their heads up and down and move their heads back and forth with the bite plate in their mouths. Analysis of these data indicated there were some problems with the NDI Wave head correction process, caused by mis-synchronization between the reference sensor and the data sensors attached to channels 9-16, which were on a secondary hardware unit. This issue affects only the MI jaw sensor, and is only a problem when there is relatively high velocity head motion so that the time lag creates inaccurate head correction. Details of this issue are available in the EMA-MAE user manual.

3.3.2 Bite-plate correction

Since the articulatory data in the head-corrected space is only roughly oriented to the subject, a key initial problem in data-processing is to calibrate the data in a more accurate way so that kinematic data is represented in a baseline articulatory working space with clear anatomical reference points and orientation (Westbury, 1991). To do this, subject calibration data is typically used to re-orient the space according to the subject's bite plate position. This can be accomplished in a number of different ways. In EMA-MAE corpus, a physical bite-plate with carefully placed sensors is used to identify the maxillary occlusal plane. Given the head-corrected measurement data recorded from the bite-plate, the goal is to translate and rotate the original coordinate space to create an articulatory working space such that the XY plane is the mid-sagittal plane and the XZ plane is the maxillary occlusal plane, with the origin placed at the upper central front incisor.

Although most EMA datasets currently available include a bite plate calibration in their preprocessing stage (Byrd, Browman, Goldstein, & Honorof, 1999; Gracco & Nye, 1993; Krista, 2011; Westbury, 1994b), none of them provide a detailed description and error analysis of this processing, or the underlying assumptions on which the calibration is based. In this section, we detail a mathematical derivation of this calibration process.

3.3.3 Target articulatory space

The target articulatory space is based on each subject's anatomy, as shown in Figure 3.3. The origin of the coordinate system is defined as the central point of the upper maxillary incisors. The vertical plane is defined as the mid-sagittal plane, and the

horizontal plane is defined as the maxillary occlusal plane, which is the plane of contact between the maxillary and mandibular natural teeth. Relative to these two coordinate planes, the X axis represents anterior-posterior motion, the Y axis represents superior-inferior motion, and the Z axis represents lateral motion. The mid-sagittal plane is thus given by the XY axes and the maxillary occlusal plane by the XZ axes.

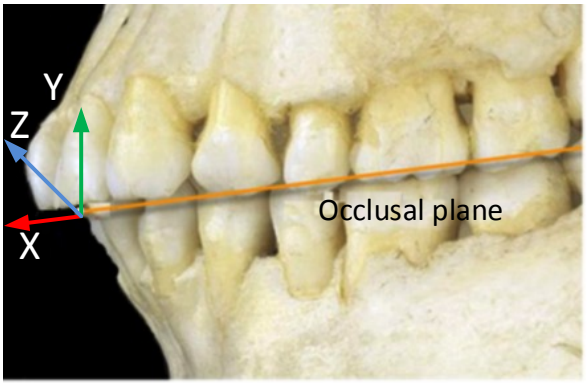

Figure 3.3 Target articulatory referenced coordinate system

By convention, the positive X axis is forward of the incisors, so that the negative X axis follows the midsagittal line of the occlusal plane toward the back of throat. The positive Z axis runs perpendicularly to the X axis on the occlusal plane toward the subject's right.. The positive Y axis is perpendicular to the occlusal plane in the upward direction.

Note that even this theoretical definition of articulatory space includes some physiological assumptions, the most prominent of which is that the midsagittal plane and

maxillary occlusal plane are in fact perpendicular. This is not all guaranteed, since the location of the temporomandibular joints are unlikely to be exactly symmetric, and even less so the detailed dental features which create the occlusal plane itself. However, these deviations are typically quite small and have minimal impact on the creation of a useful articulatory space for data analysis.

The fundamental goal of the data calibration process, called “bite-plate calibration”, is to ensure that the coordinate system represented by the data follows as closely as possible to the theoretical target articulatory space defined above.

3.3.4 Quaternion representation

The NDI Wave system uses a quaternion format representation for all orientation data. The quaternion structure is a commonly used method to represent rotation and orientation (Hart, Francis, & Kauffman, 1994) in many different fields, including computer visualization and animation, object tracking and identification, and propulsion systems due to its compactness and robustness. The quaternion format will be used in this work to represent the rotation needed to implement the optimal calibration solution, with a basic overview given here.

A quaternion is a 4-D unit vector

$$q = \left[q_0, q_x, q_y, q_z\right] \tag{3.1}$$

where $q_0^2 + q_x^2 + q_y^2 + q_z^2 = 1$. This vector can be used to represent an arbitrary single three-dimensional rotation. One of the simplest ways to visualize how a unit-normalized quaternion can be used to represent a rotation is to first consider an axis-angle viewpoint,

where a rotation is represented by an angle θ around a unit axis v. A quaternion can be thought of as a "normalized" composite axis-angle vector, given by

$$q = (\cos\frac{\theta}{2},\ \sin\frac{\theta}{2}v_x,\ \sin\frac{\theta}{2}v_y,\ \sin\frac{\theta}{2}v_z) \tag{3.2}$$

where the vector part $[q_x, q_y, q_z] = \sin(\theta/2)v$ defines the axis of rotation, and the scalar part $q_0 = \cos(\theta/2)$ defines the degree of rotation. To rotate a point, with position represented by the vector $\overrightarrow{p}$, by an angle θ around the axis v to a new position, with position p_{final},the following quaternion multiplication operation is applied:

$$p_{final} = QPQ^*, \tag{3.3}$$

where $P = [0, \overrightarrow{p}]$.

In the NDI system, sensor orientations are represented by a quaternion vector which indicates the amount of rotation a sensor has undergone relative to its established base orientation in the coordinate space. In the standard experimental configuration with a reference 6 DOF sensor REF and head-corrected data, the quaternion represents the orientation change relative to the orientation of the REF sensor plane.

3.3.5 Calibration method

Since the REF sensor is carefully placed in the midsagittal plane, and the OS and MS sensors are also carefully placed along the centerline of the bite plate, an obvious choice for calibration is to rotate the space such that these three points all lie on the XY plane, with the OS at the origin and the MS directly on the X axis. This will leave the

REF sensor in the midsaggital plane but not necessarily on the Y axis, since it may be somewhat forward or behind the vertical location of the OS sensor.

Since the distance from OS to MS, the distance from OS to REF, and the $MS - OS - REF$ angle can all be directly computed, the exact new coordinate locations for the MS and REF sensors can be easily determined. The needed rotation for calibration, for which there is a single unique solution, is thus the rotation which will rotate the $MS - OS - REF$ triangle onto these new target coordinates. Since OS is the origin in both cases, solving for this rotation focuses on the locations of the MS and REF sensors. Let $\overrightarrow{MS_L}$ and $\overrightarrow{MS_A}$ represent the MS location in local and articulatory coordinates, respectively, while $\overrightarrow{REF_L}$ and $\overrightarrow{REF_A}$ are the corresponding head reference sensor coordinates. To solve for the necessary rotation, we take the approach of solving for the set of possible rotations for the MS point and the REF point individually, then taking the intersection of the two. There are an infinite number of rotations that will rotate the original $\overrightarrow{MS_L}$ onto $\overrightarrow{MS_A}$. Figure 3.4 illustrates how to describe the set of rotation axes for this case. The bisecting vector $\overrightarrow{BS_{MS}}$ represents one possible axis, with a corresponding rotation angle of 180 degrees, and the normal vector $\overrightarrow{V_{MS}}$ represents another possible axis, with a rotation angle equal to the angle between the two points. Any line on the plane consisting of $\overrightarrow{BS_{MS}}$ and $\overrightarrow{V_{MS}}$ is also a possible axis. For any of these lines, the required rotation can be visualized as rotation along the surface of a cone, with the rotation axis as the center of the cone.

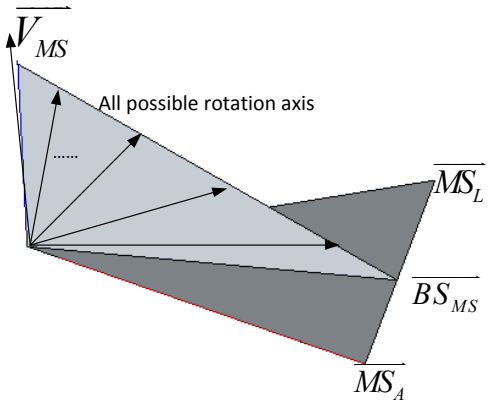

Figure 3.4 3D visualization of the possible rotation axes from $\overrightarrow{MS_L}$ onto $\overrightarrow{MS_A}$.

Mathematically, the vector normal to the plane of all possible rotation axes can be defined using the cross-product of the two axes above

$$\langle \overrightarrow{V_{MS}}, \overrightarrow{BS_{MS}} \rangle, \tag{3.4}$$

where $\overrightarrow{V_{MS}} = \langle \overrightarrow{MS_L}, \overrightarrow{MS_A} \rangle$ and $\overrightarrow{BS_{MS}} = (\overrightarrow{MS_L} + \overrightarrow{MS_A})/2$.

Similarly, there are an infinite number of rotations that will rotate the original $\overrightarrow{REF_L}$ onto $\overrightarrow{REF_A}$. By following the same steps a second plane is found that includes all possible axes which will accomplish this rotation, with normal vector $\langle \overrightarrow{V_{REF}}, \overrightarrow{BS_{REF}} \rangle$, where $\overrightarrow{V_{REF}} = \langle \overrightarrow{REF_L}, \overrightarrow{REF_A} \rangle$ and $\overrightarrow{BS_{REF}} = (\overrightarrow{REF_L} + \overrightarrow{REF_A})/2$. Solving for the intersection of these two planes gives the unique rotation axis that will simultaneously accomplish both of the desired rotations, rotating the original $MS - OS - REF$ triangle onto the XY plane:

$$\overrightarrow{Axis} = \langle \langle \overrightarrow{V_{MS}}, \overrightarrow{BS_{MS}} \rangle, \langle \overrightarrow{V_{REF}}, \overrightarrow{BS_{REF}} \rangle \rangle \tag{3.5}$$

After finding the rotation axis, it is necessary to solve for the correct rotation angle θ. Figure 3.5 illustrates this computation, based on the visualization of the rotation cone.

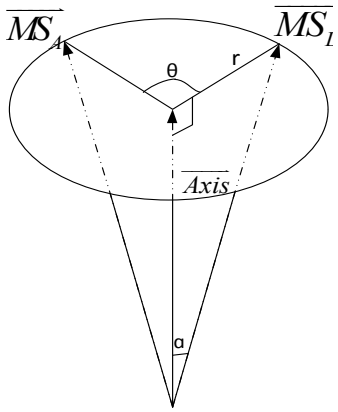

Figure 3.5 3-D visualization of the rotation angle

Using this visualization, the required rotation angle can be determined trigonometrically via

$$r = \left|\overrightarrow{MS_L}\right| \sin\alpha \tag{3.6}$$

$$\theta = 2\sin^{-1}\left(\frac{1/2\left|\overrightarrow{MS_L} - \overrightarrow{MS_A}\right|}{r}\right), \tag{3.7}$$

where α is the angle between the original MS vector MS_L and the rotation axis from this can be converted into its equivalent quaternion form

$$q_{st} = \left[\cos\left(\frac{\theta}{2}\right), \sin\left(\frac{\theta}{2}\right)\overrightarrow{Axis_x}, \sin\left(\frac{\theta}{2}\right)\overrightarrow{Axis_y}, \sin\left(\frac{\theta}{2}\right)\overrightarrow{Axis_z}\right]. \tag{3.8}$$

This q can applied using quaternion multiplication as given in Equation (3.3) to any data record to transform into the appropriate articulatory space coordinate system.

3.3.6 Palate mesh interpolation

For each subject, the palate record includes a trace of the mid-sagittal palate line, a series of transverse traces across the palate, and both inner perimeter and outer perimeter dental traces at the gum line. Together with the bite plate record, this information provides reference data that can be used to calculate physiologically-referenced vocal tract measures. The palate mesh is computed using the thin plate spline method (Yunusova et al., 2012) with a smoothing factor of 0.05 as recommended by error and variance analysis, with a vertical half-sensor offset to account for the wand sensor thickness.

3.4 Articulatory feature extraction

3.4.1 Raw EMA measurement or vocal tract feature

The EMA technique provides a simple way to measure the mechanism of articulatory motion. The measured trajectory consists of a set of position coordinates for each sensor during speech. However, the reliability and usefulness of these raw position measurements as a characterization of the speech generation process, and how they reflect the discriminability of different phonemes, are still unknown. Some prior research has suggested that raw EMA measurements are reliable cues to the acoustic signal. Toda (Toda et al., 2004) showed that the speech spectrum can be produced from EMA measurements by learning statistical dependencies between position trajectory and the

corresponding speech signal, which indicates that the raw measurement do relate to the output of the speech generation process. Most research on acoustic-to-articulatory inversion uses raw position EMA measurement, with relatively good results. While these attempts do provide indirect evidence of the relation between EMA measurements and speech production mechanism, there is still an open question about the best articulatory variables to use for both this and other tasks.

From the perspective of acoustic-to-articulatory inversion and related applications, there are several reasons that raw EMA measurement may not be the best features to use:

1. The main goal is to identify articulatory features that relate to signal acoustics. These acoustics are primarily driven by the cross-section of the vocal tract opening. Without reference to the surrounding tissue, and in particular to the upper palate which bounds the vocal tract opening, direct sensor measures are not as connected to vocal tract shape, and therefore to acoustics, as they could be.

2. The EMA measures represent only the locations of very small number of points on the vocal tract. However, the vocal tract is a very complex structure which cannot be fully characterized by such a small number of articulatory points. By incorporating additional information such as palate position, dental boundary, or inferred tongue shape using sensor orientation, it is possible to increase the amount of information contained in the articulatory features.

3. In many cases, the most acoustically relevant articulatory features may be a combination of sensor positions or sensor positions and orientations, and it

would be more effective to combine the sensor data in an appropriate way before modeling. One example of this is lip opening – while upperlip (UL), lowerlip (LL), positions (6 variables in all, including all three coordinates of each sensor) are all relevant to acoustics, the simple measure of vertical lip opening = (UL_y - LL_y), is much more efficient representation.

In order to derive reliable and phonetic meaningful features to characterize vocal tract shapes from EMA measurements, a model-based approach is used to estimate the vocal tract configuration from direct EMA measurement in this dissertation. Several theoretical models that describe the speech production process have been proposed in previous research. In this work, we will primarily use Maeda's model, which represents a mid-sagittal configuration of the vocal tract (Maeda, 1990).

3.4.2 Proposed articulatory feature

A geometric transformation from the EMA kinematic measurements to vocal tract (VT) parameters based on the mid-sagittal representation of the vocal tract from Maeda's model has been developed. These parameters include the following articulatory feature variables:

Table 3.1 Articulatory features

	Description
VT1	Tongue dorsum normalized horizontal position
VT2	Tongue dorsum vertical height to hard palate
VT3	Tongue body normalized horizontal position
VT4	Tongue body vertical height to hard palate
VT5	Tongue apex normalized horizontal position
VT6	Tongue apex vertical height to hard palate
VT7	Normalized horizontal lip protrusion
VT8	Normalized vertical lip separation

To create a normalized working space, the distance between the center incisors and the middle point of the back molar from each speaker's bite plate record is used as a normalization scalar when calculating the horizontal position of the tongue, to give better information regarding tongue position relative to the whole vocal tract across individuals. The horizontal (X axis) variables VT1, 3, 5, and 7, are all calculated directly from sensor position divided by this normalization constant. The hypothesis is that this will lead to improvement in cross-subject variability but not variability or inversion accuracy within a single subject. The vertical (Y axis) variables VT2, 4, and 6; however, are computed from the vertical distance between the sensor position and the palate, representing vocal tract height at the sensor positions, including two midsagittal positions and one lateral position. It is hypothesized that these vertical articulatory variables will be significantly more representative of vocal tract height and cross section area and therefore of acoustic

spectral characteristics both within and across subjects. Lip protrusion VT7 is taken directly from the sensor X position without any normalization, and vertical lip separation VT8 is calculated as lip separation rescaled to a [0, 1] working space.

$$VT8 = \frac{(UL_y - LL_y) - (UL_y - LL_y)_{closed\ position}}{(UL_y - LL_y)_{max}} \tag{3.9}$$

3.4.3 Working space analysis

To compare the working space based on direct EMA measurements with that using the proposed palate-referenced features, the variance of the features is used, overall and within specific vowel configurations. An emphasis is placed on the variance in the vertical direction where the palate referencing has significant impact on the feature information. Figure 3.6 compares the feature spaces for the vowel ***/i:/*** *(in word "s**ee**, h**ea**t")* for a female native English speaker. Focusing on the vertical dimension, it can be seen that the overall working space is smaller and more compressed in the proposed palate-referenced feature space.

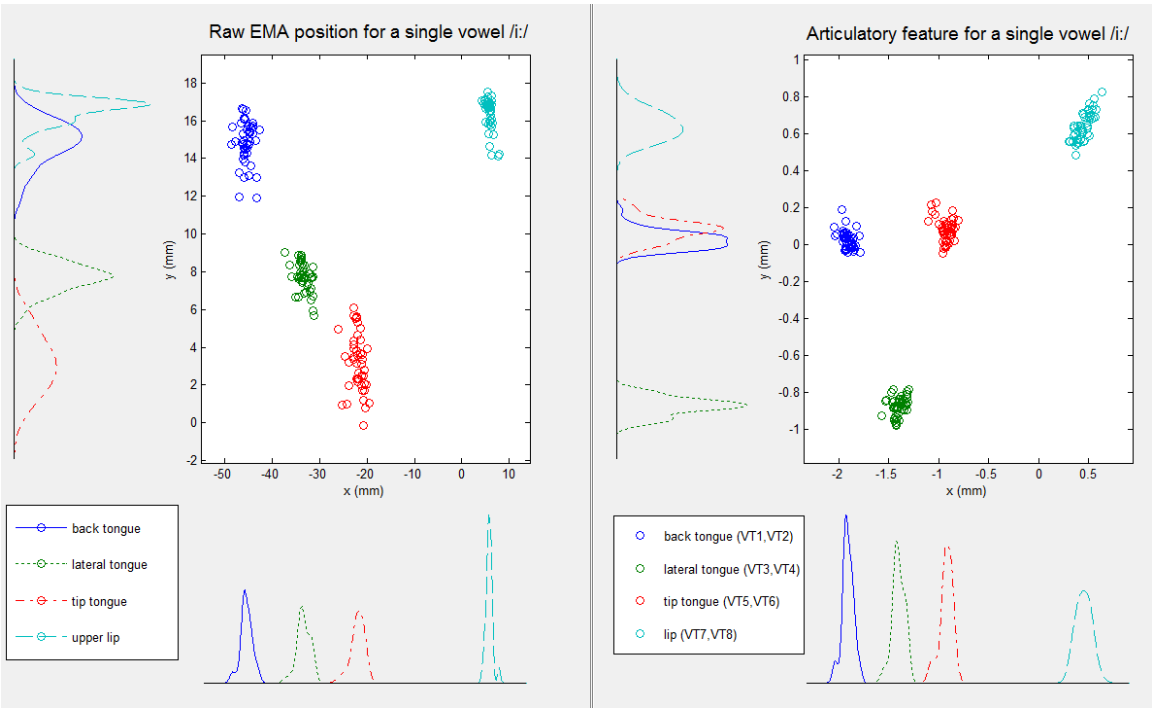

Figure 3.6 Feature space of vowel ***/i:/*** for direct sensor measures (left) and proposed articulatory features (right)

To quantify the difference between these features spaces, an ANOVA analysis for the vertical direction features is implemented across all 20 native English speakers for the vowel ***/i:/***, with results shown in Figure 3.7. The proposed features have a lower F score and higher p value compared to the raw sensor movement, indicating a lower cross-speaker variance for this vowel in the articulatory feature working space. By reducing cross-speaker variance, the proposed articulatory features reduced individuality and represent a more common working space across speakers.

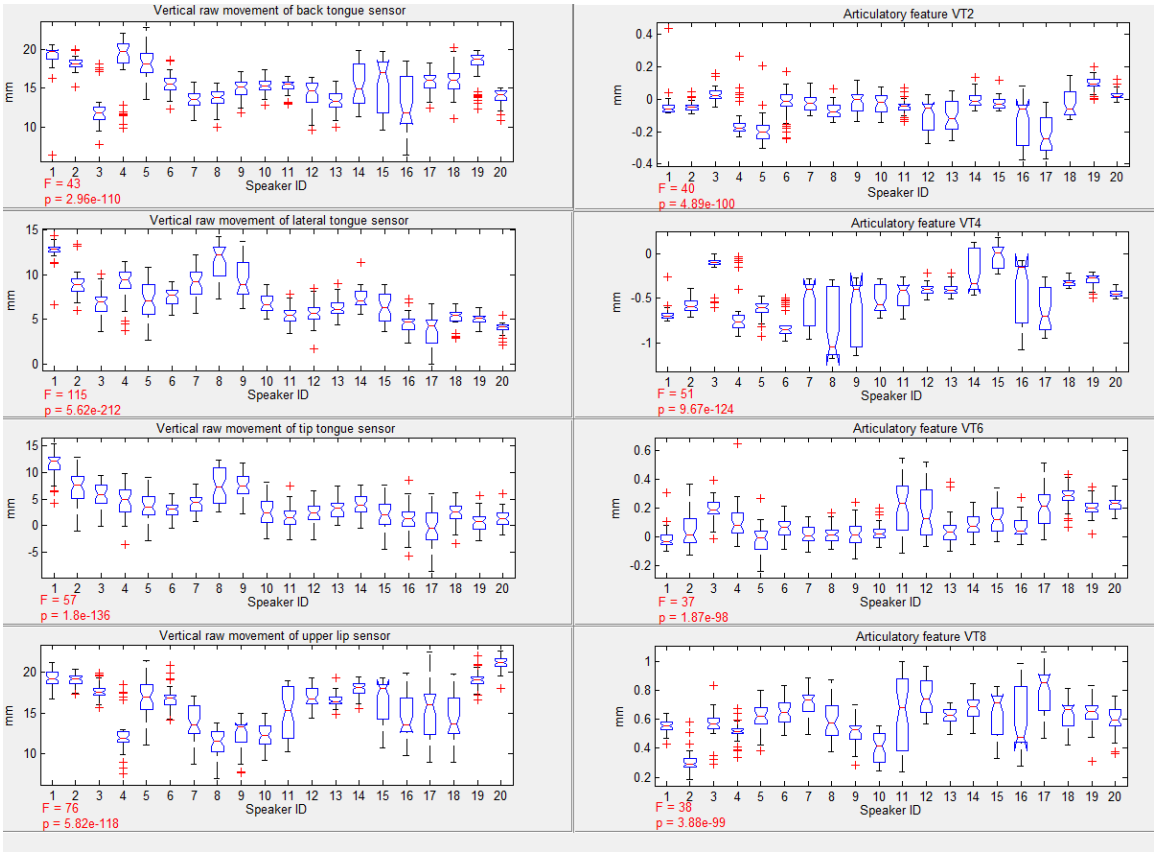

Figure 3.7 ANOVA analysis of single vowel ***/i:/*** across speakers

To illustrate inter-vowel difference and feature discriminability, Figure 3.8 compares the working spaces for three different vowels for a single female speaker, while Figure 3.9 shows the corresponding ANOVA analysis using the combined data from all 20 speakers. The selected vowels are ***/i:/*** (as in “heat”), ***/ou/*** (as in “home”), and ***/ei/*** (as in “ate”), which are acoustically distinct and widely separated in terms of formant values. It can be seen that the overlap between the vowels is significantly reduced using the proposed articulatory variables. The larger F score shows that the separability between different vowels is higher for the proposed features than for the raw sensor movements. This

supports the hypothesis that the proposed features do a more effective job of representing articulatory motion for tasks such as speech recognition and modeling.

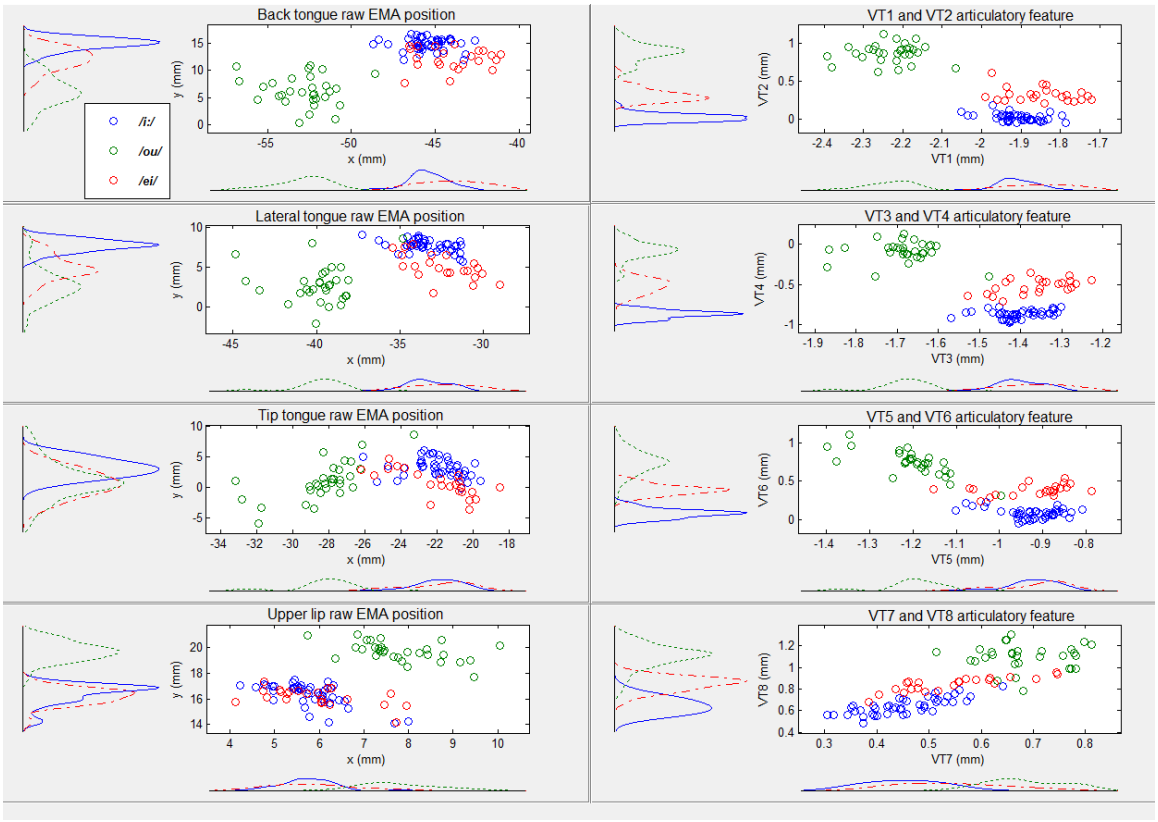

Figure 3.8 Feature space distributions for ***/i:/*** , ***/ou/*** and ***/ei/*** for direct sensor measures (left) and proposed articulatory features (right)

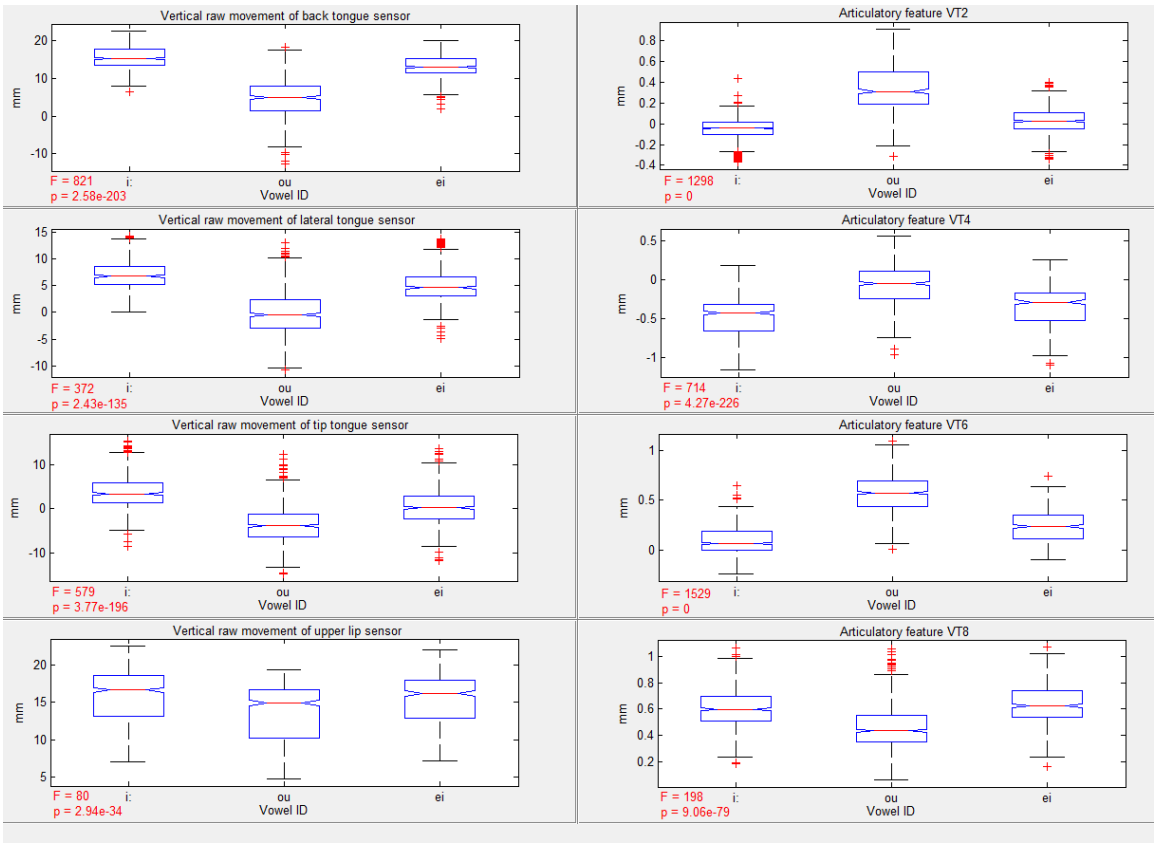

Figure 3.9 Feature space ANOVA analysis vowels ***/i:/, /ou/,*** and ***/ei/***, using combined data from all 20 speakers

The results in figures 3.6 and 3.7 show that cross-speaker variance is significantly reduced in the palate-reference feature space, and figures 3.8 and 3.9 show that these features have more discriminability between distinct vowels, strongly suggesting that the new features have better discriminatory representations than direct kinematic data. These observations suggest that the proposed palate-referenced articulatory features are a more effective overall representation, with a more compact working space and better discrimination ability.

3.5 Summary

This chapter has introduced the Marquette EMA-MAE bilingual corpus. This is the first EMA dataset to have such a substantial speaker set, including 40 speakers representing two native language groups, and also the first to include lateral tongue and lip sensors for additional 3-dimensional characterization of tongue shape and lip rounding. Data preprocessing methods for this dataset includes head motion correction, bite plate calibration and palate surface estimation, all of which have been discussed here in detail. For the purpose of acoustic-to-articulatory inversion, we have proposed an articulatory feature extraction method using EMA position measurements based on Maeda's vocal tract model. A direct comparison of the working space showed that the proposed articulatory features have smaller within-vowel variance and more discriminative ability across vowels within the same speaker. In the next Chapter, we will use the Marquette EMA-MAE corpus to evaluate the proposed acoustic-to-articulatory inversion system and will see the performance of the proposed articulatory features under this new inversion system.

4 Acoustic-to-articulatory inversion system

4.1 Introduction

This chapter describes a Hidden Markov Model based mapping that estimates articulatory parameters from an acoustic speech signal. Unlike the HMM based inversion models discussed in Chapter 2, this dual model maps the acoustic and articulatory domains through state sequence alignment, in the context of a conventional HMM. The acoustic and articulatory features are treated as two streams in the training stage in order to ensure that the acoustic and articulatory HMMs have matching state boundaries. The parameters of these two HMMs are independently estimated, and no correlation between the acoustic and articulatory transition variables are taken into account. The performance is evaluated by root mean square error as well as correlation between estimated and true articulatory parameters. Sections 4.2 and 4.3 will discuss the complex nature of the inversion problem and the basic framework of the HMM inversion system. Experimental set-up and results under different model parameter configurations will be given in section 4.4 and 4.5, respectively. Finally, section 4.6 summarizes the HMM inversion model used in this dissertation.

4.2 The nature of acoustic-to-articulatory inversion

Given that the mapping from acoustics to articulatory shape is one-to-many, as described in Chapter 2, how frequently and to what extent does non-uniqueness occur in normal human speech, and how does this affect inversion algorithms which necessarily do one-to-one inversion? Qin (Qin & Carreira-Perpinan, 2007) and his group investigated

this and found that only 5% of acoustic features mapped to a multi-modal cluster in articulatory space. This study suggests that non-uniqueness is a less frequent event and that most of the time a unique vocal tract shape is adopted for human speech production in practice.

Another issue that needs to be considered is the difference in feature complexity between the cepstrum feature used for acoustic signals and the articulatory positional features, which have a smooth, slow-varying nature. Inversion algorithms need to be able to generate less complex articulatory features. Some research has a post-processing step to smooth the inversion output, such as a low-pass filter (Richmond, 2002) or Kalman filter. In an HMM based inversion model, the differences in feature complexity are usually represented by assigning different number of Gaussian mixtures.

Unique inversion results from an acoustic speech signal are not guaranteed without imposing additional constraints. Not all configurations generated by a typical inversion model are physiologically possible in human speech production (Richmond, 2002).

4.3 HMM-based acoustic-to-articulatory inversion

Due to the ill-posed nature of the inversion problem, it is reasonable to connect the articulatory and acoustic domains through a highly abstracted phoneme level representation, instead of seeking a direct mapping, as discussed previously in Section 2.5.5. The diagram of such an acoustic-articulatory model is illustrated in figure 4.1. In this approach, the idea is to build two separate HMM in both acoustic and articulatory space through state sequence synchronization. Parallel acoustic and articulatory data is

used to train acoustic and articulatory HMMs separately. The two HMMs are aligned by state sequences for a given phonetic unit. Within each state, a GMM is used for modeling the statistical distribution of the feature vectors in each domain. The number of mixtures differs because the acoustic features have a more complex distribution than the articulatory trajectory. In the inversion stage, the test speech signal is input to the acoustic HMM to derive an optimal HMM state sequence using the Viterbi algorithm, and the corresponding aligned articulatory HMMs can be used to recover the articulatory trajectory. The articulatory HMM generates a smoothed position trajectory, using the articulatory means combined with a dynamic smooth window of the articulatory distribution, based on the maximum likelihood parameter generation algorithm described in the following section.

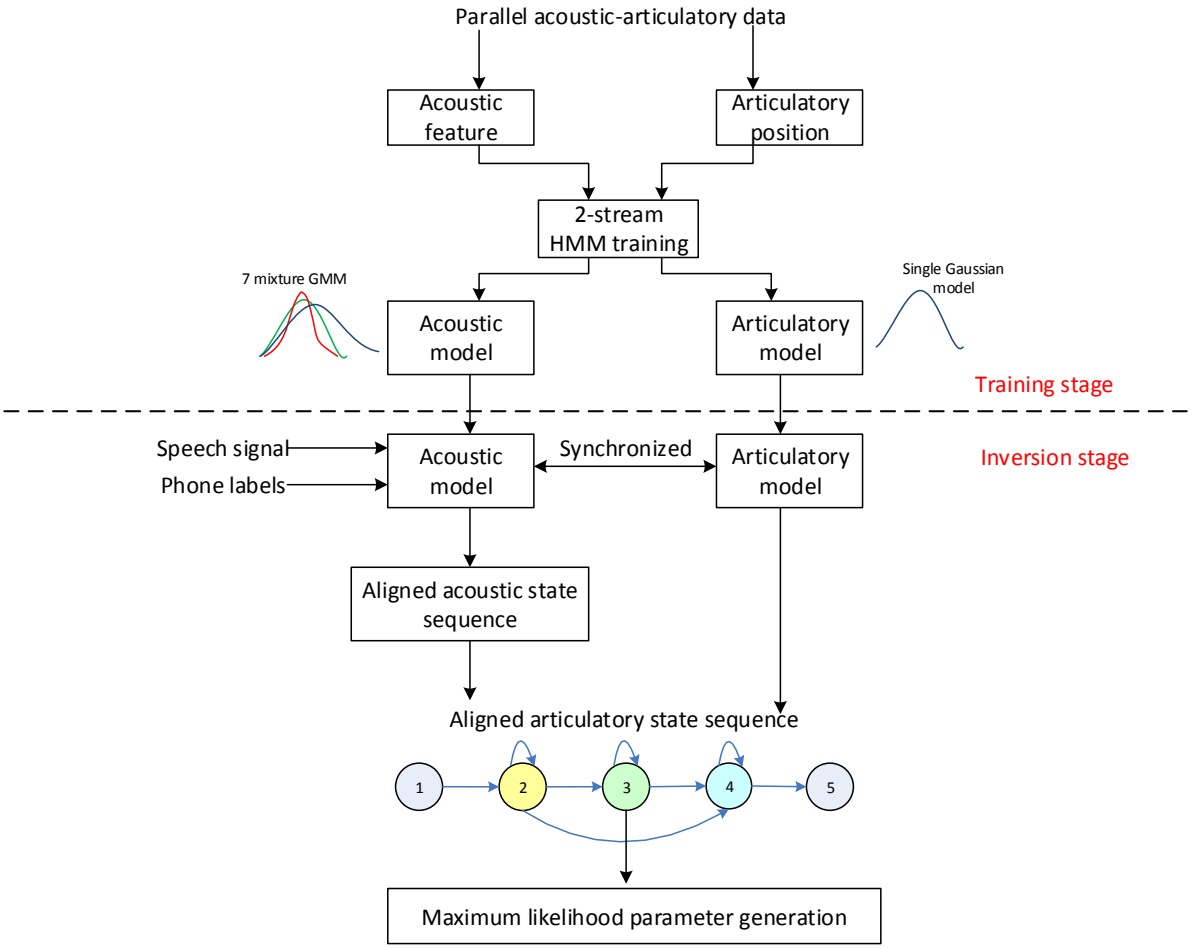

Figure 4.1 Diagram of the HMM-based articulatory-to-acoustic inversion system.

4.3.1 Training

The acoustic and articulatory HMMs are trained separately using the maximum likelihood Expectation Maximization algorithm under standard HMM training procedures. The acoustic HMM is trained first, after which the trained acoustic models are used to derive state level alignment for training of articulatory HMM parameters.

4.3.2 Forced alignment

In the inversion stage, the speech signal and phone labels are input into the acoustic HMM, and a state sequence is produced by applying forced alignment with the Viterbi algorithm. The articulatory states matching the corresponding acoustic states are concatenated into an articulatory state sequence.

4.3.3 Maximum likelihood parameter generation using dynamic features

Once the articulatory states alignment is generated, the recovery algorithm needs to estimate a smooth and slow changing articulatory trajectory from the HMM state sequence. The observation data sequence O is estimated by maximizing $P(O|Q,\lambda)$ with respect to O for a fixed state sequence $Q = [q_1,\ q_2, \dots, q_t]$. The logarithm of $P(O|Q,\lambda)$ can be written as

$$logP(O|Q,\lambda) = -\frac{1}{2}O^T\Sigma^{-1}O + O^T\Sigma^{-1}U + K, \tag{4.1}$$

where

$$\Sigma^{-1} = diag[\Sigma^{-1}_{q_1,i_1}, \Sigma^{-1}_{q_2,i_2}, \dots, \Sigma^{-1}_{q_t,i_t}] \tag{4.2}$$

and

$$U = [\mu^T_{q_1,i_1}, \mu^T_{q_2,i_2}, \dots, \mu^T_{q_t,i_t}]^T. \tag{4.3}$$

Here μ_{q_t,i_t} and Σ_{q_t,i_t} are the $3M\times1$ mean vector and the $3M\times3M$ covariance matrix associated with i-th mixture of state q_t, respectively. The constant K is independent of O.

It is clear from this that $P(O|Q,\lambda)$ is maximized when $O = U$, that is, when the output parameter vector sequence is the sequence of mean vectors, resulting in a step-wise function, because this is the maximum likelihood sequence for the sequence of Gaussians. In order to recover a smooth articulatory trajectory from articulatory HMM model parameters, dynamic features can be used, as described in (Tokuda, Yoshimura, Masuko, & Kobayashi, 2000). The basic idea is to build a matrix which includes dynamic features and use this information to smooth the output state mean value. The transformation is given as

$$o_t = w_t c_t \tag{4.4}$$

$$o_t = [c_t, \Delta c_t, \Delta^2 c_t\ ]\ , \tag{4.5}$$

where o_t is the feature vector at time t,which includes static features c_t , dynamic delta (velocity) coefficients Δc_t, and delta-delta (acceleration) coefficients $\Delta^2 c_t$. w_t is a 3 by T transformation matrix, where *T* is the total number of frames, defined by

$$w_t = \begin{bmatrix} 0, & \dots, & 0 & 0, & \dots & 1, & \dots, & 0, & 0, & \dots, & 0 \\ 0, & \dots, & 0 & w^1(-L), & \dots, & w^1(0), & \dots, & w^1(L), & 0, & \dots, & 0 \\ 0, & \dots, & 0 & w^2(-L), & \dots, & w^2(0), & \dots, & w^2(L), & 0, & \dots, & 0 \end{bmatrix}. \tag{4.6}$$

The elements in the first row are all zero except for the t^{th} column which corresponds to the static feature at the t^{th} frame. The second and the third rows represent the coefficients for computing the dynamic delta and delta-delta features. *L* is the window length defined to calculate those features. The augmented feature vector over t frames can be written as follows:

$$\begin{bmatrix} o_1 \\ o_2 \\ \vdots \\ o_t \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_t \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_t \end{bmatrix}.. \tag{4.7}$$

Letting $\mathrm{W} = \begin{bmatrix} \mathrm{w}_1 \\ \mathrm{w}_2 \\ \vdots \\ \mathrm{w}_t \end{bmatrix}$, we have

$$O = WC. \tag{4.8}$$

Maximizing $P(O|Q,\lambda)$ with respect to O is equivalent to maximizing with respect to C. By setting

$$\frac{\partial log P(WC|Q,\lambda)}{\partial C} = 0, \tag{4.9}$$

we obtain a set of equations

$$W^T\Sigma^{-1}WC = W^T\Sigma^{-1}U^T. \tag{4.10}$$

Solving these equations, the static feature trajectory estimate is recovered from the state sequence parameters via

$$C = W^{-1}\Sigma(W^T)^{-1}W^T\Sigma^{-1}U^T . \tag{4.11}$$

Figure 4.2 shows an example of a recovered trajectory with dynamic features. Unlike the stepwise mean output of a conventional HMM, the output from this model is a smoothed trajectory.

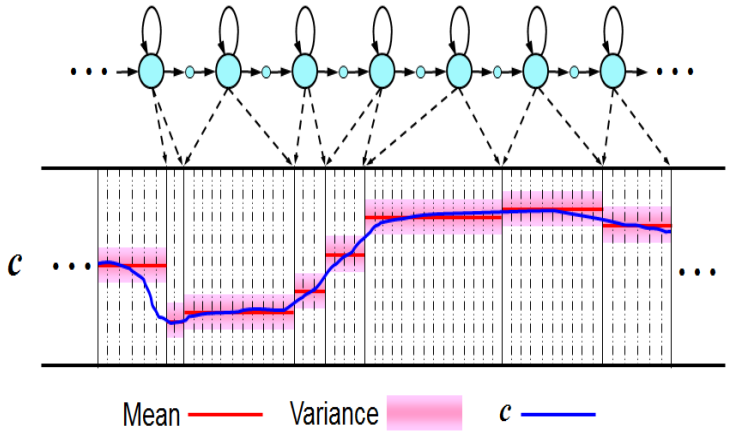

Figure 4.2 Recovered static feature incorporating dynamic features

4.4 Experimental set up

4.4.1 Data pre-processing

A male native English speaker's acoustic and kinematic data from the EMA-MAE database has been used for these experiments. The acoustic feature vector has 39 dimensions including 12 MFCCs plus energy, along with their first and second derivatives. The EMA data is decimated (down sampled with an anti-aliasing filter) to 100 Hz to match the 10 ms frame shifting rate of the acoustic features. Five state left-to-right mono-phone models with differing number of Gaussian mixtures per state are used for training and testing. A 9-fold cross validation test was chosen to measure the accuracy of the inversion, selected for convenience since there are 198 utterances, an even multiple of 9. These utterances are divided into 9 partitions consisting of 22 sentences, with one partition used for the testing and the other 8 partitions used for training, and the process repeated 9 times with each partition used as test data once.

4.4.2 Evaluation metrics

Metrics for performance evaluation include the deviation between the actual and estimated articulatory position values and the correlation with the actual articulator trajectories. Denoting the actual values of the articulator measure as y and the corresponding values of the estimated output as $f(x)$, the normalized RMS error over the whole test set is calculated as:

$$E_{rms} = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2}}{std(y)}, \tag{4.12}$$

where m is the number of examples in the test set. y_i is the true articulatory variable value, $f(x_i)$ is the inversion output, and $std(y)$ is the standard deviation of the articulatory variable across the full test set. This normalized RMS error is used to evaluate inversion performance for the proposed articulatory feature across different scales.

Correlation is

$$r = \frac{\sum_{i=1}^{m}(f(x_i)-\overline{f(x_i)})(y_i-\overline{y_i})}{\sqrt{\sum_{i=1}^{m}(f(x_i)-\overline{f(x_i)})^2 \sum_{i=1}^{m}(y_i-\overline{y_i})^2}}, \tag{4.13}$$

where $\overline{f(x_i)}$ and $\overline{y_i}$ are the means of the estimated and actual articulatory values, respectively.

A good articulatory inversion system is expected to obtain low RMS error and high correlation with respect to real articulatory data. In prior work, several different EMA datasets have been used across various different methodologies, which makes it

difficult to compare results or have a strong frame of reference for expected performance. However, MOCHA-TIMIT has been the most widely used EMA dataset. The lowest RMS error reported is from Richmond's trajectory mixture density networks (Richmond, 2002) which is 0.99mm on the MNGU0 speaker data.

4.5 Results

4.5.1 Model complexity influence in terms of state alignment

The quality of the HMM state alignment, both for training the articulatory HMMs and for deriving the articulatory feature inversion, is important to the overall performance. In the HMM based inversion described in section 4.3, the HMM state alignment is derived from acoustic HMMs by forced alignment with the phone label sequence. This section provides a closer examination of how the quality of these alignments impact the inversion performance.

The accuracy of the derived HMM alignment depends on the quality of acoustic models. In conventional speech recognition Gaussian mixture models are used to model the state emission distribution. A higher number of mixtures normally yields better acoustic models given sufficient training data. Thus increasing the number of mixtures in the acoustic HMM can improve the quality of state alignment for articulatory HMM training and inversion. In this experiment, the number of mixtures is increased from 1 up to 12, and the inversion performance is compared under these different alignments. The average normalized RMS error and correlation are used to analyze the effect of using

different number of Gaussian mixture components in the acoustic model. Figure 4.3 and table 4.1 - 4.2 show the results.

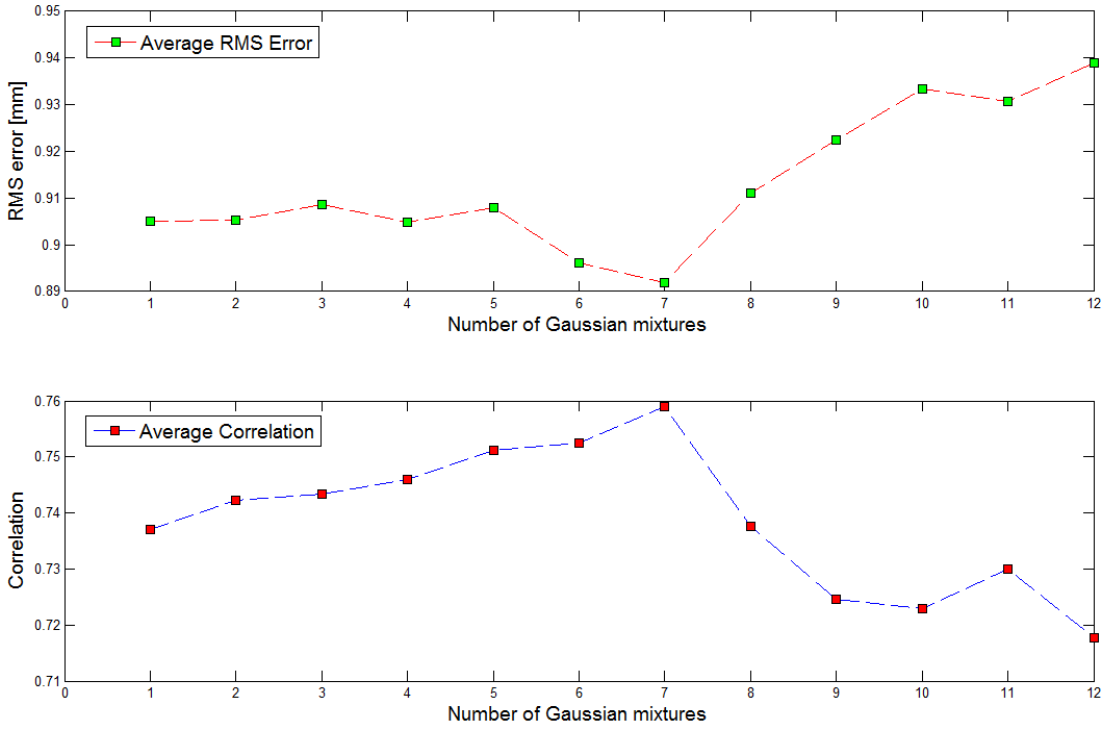

Figure 4.3 Inversion performance for an increasing number of acoustic mixtures

Figure 4.3 shows the average normalized RMS error and correlation across eight articulatory features under different acoustic models. The higher the number of mixture components used for alignment, the lower the normalized RMS error and the higher the correlation initially. The best performance is observed at seven mixtures, but from 8 to 12 mixtures, the inversion performance drops back down to the level of single mixture. Normally, the upper limit on the number of mixtures, which is directly proportional to the total number of model parameters, is determined by the quantity of training data. In order

to ensure that the model is sufficiently trained and results will be generalizable to new unseen data, a sufficient number of examples is required to estimate means and variances for each mixture in each state. If the number of parameters is increased beyond this point, the model will begin to over fit to the training data, and test set accuracy will begin to decrease. In this case, using more than 8 mixtures indicates that the model is starting to over fit the training data.

Tables 4.1 and 4.2 show the normalized RMS error and correlation for individual articulatory features under different numbers of mixtures. The best performance can be found using 7 mixtures for every articulatory feature which indicates that the alignment under this model is optimal in terms of inversion accuracy.

Table 4.1 Normalized RMS error for individual articulatory features

Articulatory feature / Number of mixtures	VT1	VT2	VT3	VT4	VT5	VT6	VT7	VT8
1	0.94	0.90	1.03	0.96	1.02	0.77	0.79	0.87
2	0.96	0.90	1.01	0.95	1.00	0.77	0.80	0.87
3	0.92	0.90	1.02	0.96	1.02	0.78	0.80	0.85
4	0.92	0.89	1.01	0.95	1.02	0.78	0.82	0.86
5	0.95	0.90	1.00	0.96	1.00	0.77	0.79	0.86
6	0.93	0.89	1.00	0.94	1.00	0.76	0.79	0.87
7	0.92	0.87	0.99	0.94	0.99	0.75	0.79	0.87
8	0.93	0.91	1.00	0.98	1.00	0.78	0.81	0.87
9	0.94	0.91	1.00	0.97	1.00	0.77	0.81	0.87
10	0.95	0.90	0.99	0.98	1.00	0.79	0.80	0.87
11	0.95	0.92	1.00	0.97	1.03	0.80	0.81	0.88
12	0.97	0.92	1.01	0.98	1.03	0.79	0.81	0.88

Table 4.2 Correlation for individual articulatory features

Articulatory feature / Number of mixtures	VT1	VT2	VT3	VT4	VT5	VT6	VT7	VT8
1	0.71	0.76	0.70	0.72	0.69	0.78	0.73	0.81
2	0.71	0.77	0.69	0.71	0.69	0.79	0.77	0.80
3	0.73	0.77	0.69	0.71	0.69	0.79	0.78	0.81
4	0.72	0.76	0.71	0.72	0.69	0.80	0.75	0.81
5	0.73	0.76	0.71	0.73	0.69	0.81	0.76	0.81
6	0.74	0.76	0.71	0.73	0.69	0.81	0.76	0.82
7	0.74	0.78	0.72	0.74	0.70	0.82	0.79	0.83
8	0.73	0.74	0.71	0.70	0.70	0.79	0.75	0.80
9	0.73	0.74	0.71	0.69	0.69	0.79	0.74	0.79
10	0.72	0.74	0.71	0.70	0.69	0.79	0.74	0.80
11	0.72	0.74	0.71	0.69	0.70	0.79	0.74	0.79
12	0.71	0.73	0.71	0.70	0.69	0.79	0.73	0.80

Overall, these results in general agree with the idea that even when the articulatory HMM uses a single Gaussian to generate output, the inversion system can still benefit from using a more complex acoustic HMM to derive better state alignment.

4.5.2 Dynamic window impact

The maximum likelihood parameter generation algorithm described in 4.3.3 uses static and dynamic features to recover the slowly changing trajectory. The coefficients in the W matrix are the same coefficients used to calculate delta and delta-delta features. Different window types will have different impact on the recovered trajectory. In this section, the impact of two different common windows on the inversion performance is investigated. Normally, the delta coefficient (velocity) is an MSE estimate of the slope of a line passing the data points. The solution is derived from linear regression by the given

data. Delta-delta coefficients (acceleration) are traditionally calculated as the delta of the delta. However, the best estimate of the acceleration in a maximum likelihood sense is the high order coefficient of a second order polynomial passing through the data. In the following experiments, two different methods to calculate the velocity and acceleration coefficients are implemented for the inversion system:

Method #1: Analytic solution of the velocity from the first order regression, approximate estimation of acceleration from repeated first order regression on the velocity coefficients. (HTK method). In automatic speech recognition, such as in the well-known Hidden Markov Model Toolkit (HTK) it has traditionally been common to calculate delta and delta-delta coefficients as follows:

$$\Delta c_t = \frac{\sum_{\theta=-n}^{n} \theta c_{t+\theta}}{\sum_{\theta=-n}^{n} \theta^2} \tag{4.14}$$

$$\Delta^2 c_t = \frac{\sum_{\theta=-n}^{n} \theta \Delta c_{t+\theta}}{\sum_{\theta=-n}^{n} \theta^2}, \tag{4.15}$$

where c_t is the static feature at frame t, and n is the half window length used to calculate dynamic feature at frame t. Choosing $n = 1$, which is a 3-frame window for calculating velocity and a 5-frame window for calculating acceleration at frame t, we have

$$\Delta c_t = -0.5 c_{t-1} + 0.0 c_t + 0.5 c_{t+1} \tag{4.16}$$

$$\Delta^2 c_t = 0.25 c_{t-2} - 0.5 c_t + 0.25 c_{t+2} \, . \tag{4.17}$$

For an n of 2, which is a 5-frame window for calculating velocity and a 9-frame window for calculating acceleration, at frame t, we have

$$\Delta c_t = -0.2 c_{t-2} - 0.1 c_{t-1} + 0.0 c_t + 0.1 c_{t+1} + 0.2 c_{t+2} \tag{4.18}$$

$$\Delta^2 c_t = 0.04c_{t-4} + 0.04c_{t-3} + 0.01c_{t-2} - 0.04c_{t-1} - 0.1c_t -$$

$$0.04c_{t+1} + 0.01c_{t+2} + 0.04c_{t+3} + 0.04c_{t+4} \ . \quad (4.19)$$

This type of window is denoted as W1_3_5 (window type 1, 3 points for velocity, 5 points for acceleration) and W1_5_9 (window type 1, 5 points for velocity, 9 points for acceleration)

Method #2: Analytic solution for the velocity and acceleration coefficients from the first and second order regression analysis (HTS method).

Theoretically, the analytic solution of acceleration coefficients should be estimated from a second order polynomial rather than applying the linear regression to the delta/velocity coefficients. The HMM based speech synthesis system HTS uses the analytic solution to calculate dynamic coefficients as follows:

$$\Delta c_t = \frac{\sum_{\theta=-n}^{n} \theta c_{t+\theta}}{\sum_{\theta=-n}^{n} \theta^2} \quad (4.20)$$

$$\Delta^2 c_t = 2\frac{\sum_{\theta=-n}^{n} \theta^2 c_{t+\theta} - \frac{1}{N}(\sum_{\theta=-n}^{n} \theta^2)(\sum_{\theta=-n}^{n} c_{t+\theta})}{\sum_{\theta=-n}^{n} \theta^4 - \frac{1}{N}\sum_{\theta=-n}^{n} \theta^2}, \quad (4.21)$$

where $N = 2n + 1$ is the width of the window used to calculate dynamic features at frame t. For $n = 1$, which is a 3-frame window for calculating both velocity and acceleration coefficients, for frame t, we have

$$\Delta c_t = -0.5c_{t-1} + 0.0c_t + 0.5c_{t+1} \quad (4.22)$$

$$\Delta^2 c_t = 0.5c_{t-1} - c_t + 0.5c_{t+1} \ . \quad (4.23)$$

For n is 2, which is a 5-frame window for calculating velocity acceleration coefficients at frame t, we have

$$\Delta c_t = -0.2c_{t-2} - 0.1c_{t-1} + 0.0c_t + 0.1c_{t+1} + 0.2c_{t+2} \quad (4.24)$$

$$\Delta^2 c_t = 0.125c_{t-2} - 0.0625c_{t-1} - 0.125c_t - 0.0625c_{t+1} + 0.125c_{t+2} . \quad (4.25)$$

This type of window is denoted by W2_3_3 (window type 2, 3 points for velocity, 3 points for acceleration) and W2_5_5 (window type 2, 5 points for velocity, 5 points for acceleration)

In order to compare the two different methods under the same window length for both velocity and acceleration, W2_3_5 and W2_5_9 are implemented to match W1_3_5 and W1_5_9.

It should be noted that these two methods use the same computation for delta/velocity coefficients, and only differ in terms of the formula used for delta-delta/acceleration. To analyze the impact of inversion performance as a result of window type, the inversion results are compared across the two methods. The average normalized RMS error and correlation are given in table 4.3:

Table 4.3 Inversion results comparison

Inversion result / Window type	Average normalized RMS error	Average correlation
W1_3_5	0.90	0.74
W1_5_9	1.09	0.67
W2_3_5	0.90	0.74
W2_5_9	1.02	0.68
W2_3_3	0.90	0.74
W2_5_5	0.99	0.70

The inversion performance is better when using the window coefficients computed from Method #2. This might suggest that the correct analytic solution is the optimal window for inversion.

From this table, it can also be seen that shorter windows within the same window type give better inversion performance. By looking at the recovered trajectory in figure 4.4, we see empirically that a larger window generates a noisy inversion result, while a smaller one gives smoother output. The better performance using a smaller window suggests that the feature dynamics are fast enough that the large window is over smoothing, therefore inaccurately calculating the second order term. Thus it is not always true that a longer dynamic window is capable of capturing longer range correlations between frames and should generate smoother output trajectories. A larger window is only theoretically better if the acceleration is not changing very fast.

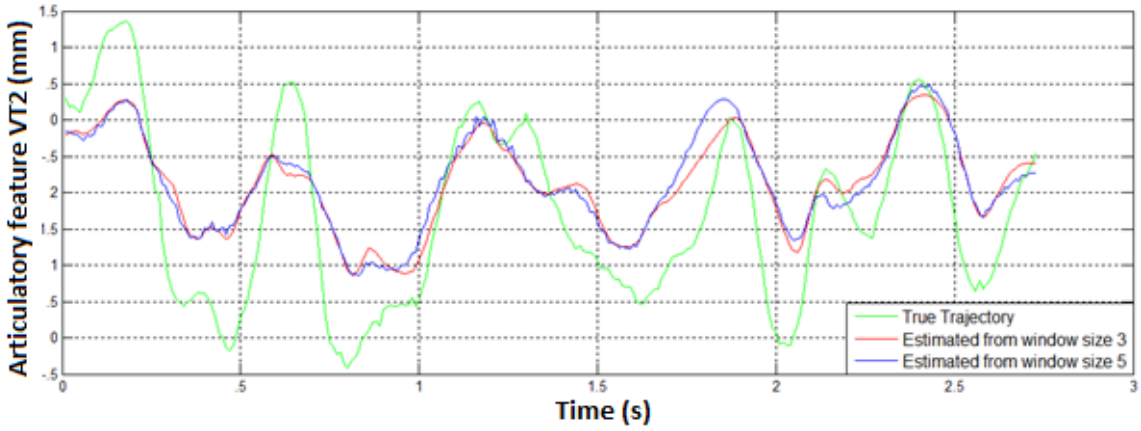

Figure 4.4: Estimated trajectory for the VT2 articulatory feature of a test utterance: "The object of the game was to produce a good time". The output from window size 3 (red line) is smoother than that of window size 5 (blue line)

4.5.3 Articulatory feature .VS. Direct sensor movement

The selection of effective articulatory features is an important component of acoustic-to-articulator inversion. Although it is common to use direct articulatory sensor measurements as feature variables, this approach fails to incorporate important physiological information such as palate height and shape and thus is not as representative of vocal tract cross section and the associated acoustics. In this experiment, we use the HMM inversion system to compare two sets of articulatory parameters. The first is the direct sensor position, which is the typical articulatory feature variable used for most studies of articulatory kinematics and acoustic-to-articulatory inversion. The second is the set of articulator features described previously in Chapter 3. The hypothesis is that the proposed articulatory features should give better inversion performance because these features are palate referenced and normalized with respect to the articulatory working space, and therefore a better representation of the vocal tract. The quality of the feature

representation is evaluated quantitatively through measurement of acoustic-to-articulator inversion error.

In this experiment, the HMM inversion system is used to estimate the articulatory parameters from the acoustic signal. The experimental set up is the same as previously described. Two sets of articulatory feature vectors are implemented, the first being the direct x and y position values of the designated EMA sensors, and the second being the proposed articulatory features in chapter 3, in both cases with their first and second derivatives.

Figure 4.5 illustrates the measured and reconstructed time trajectories of raw sensor coordinates and vertical articulatory feature for a test utterance. The vertical features are chosen for illustration because these distances between tongue sensors and palate surface which best represent the cross section of the vocal tract and lip opening, whereas palate reference for horizontal features is likely to have much less impact.

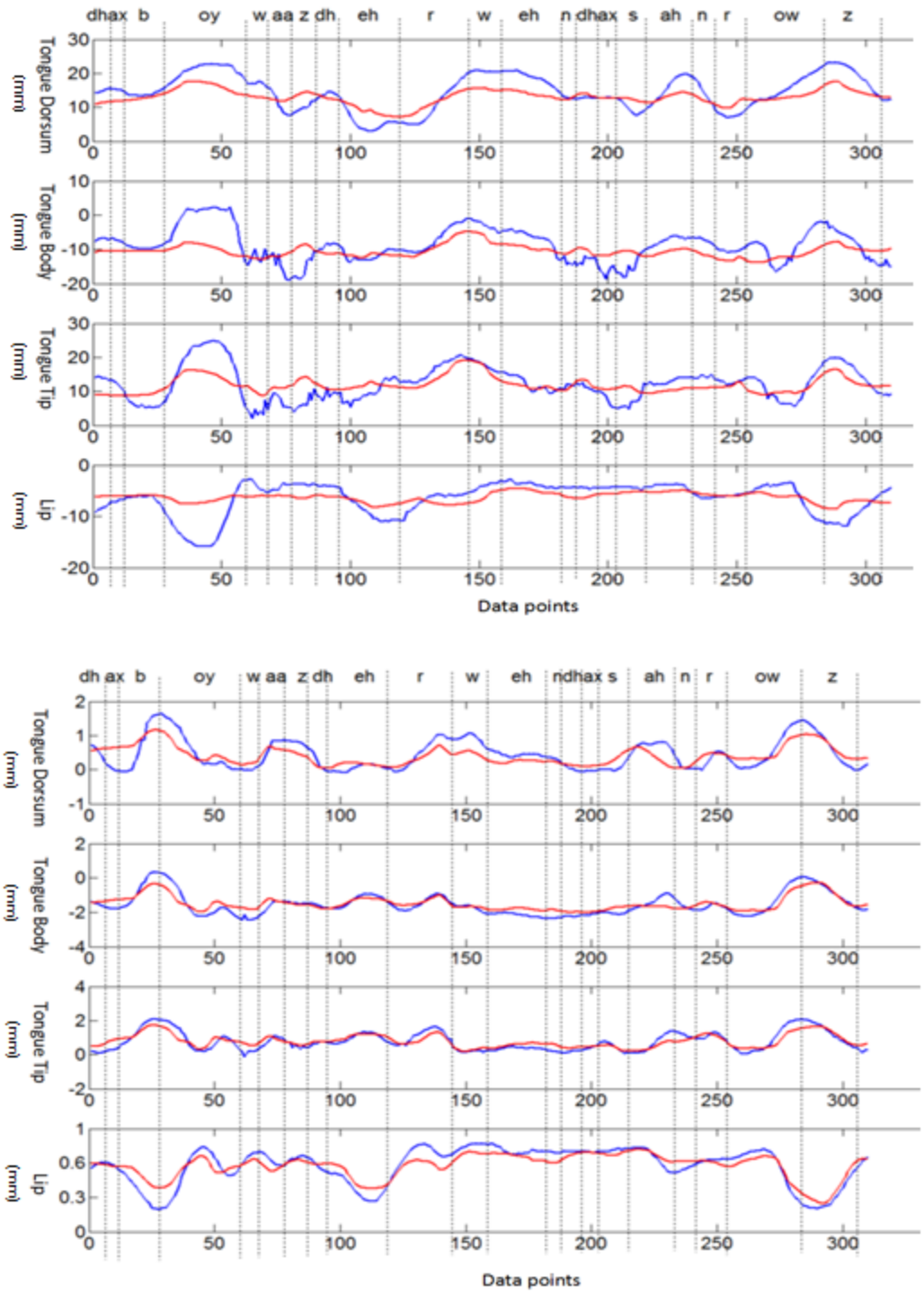

Figure 4.4 Measured (blue lines) and reconstructed (red lines) trajectories of the direct measures (upper) and articulatory features (lower), in the test sentence "The boy was there when the sun rose". Phone boundaries are shown by vertical bars

Table 4.4 Normalized RMS error and correlation coefficients between acoustic-to-articulator inversion estimates and actual trajectories.

	Normalized RMS error		Correlation	
	Sensor space	AF space	Sensor space	AF space
Dorsum	1.20	0.93	0.66	0.72
Body	1.45	0.92	0.62	0.73
Tip	1.31	0.98	0.60	0.79
Lips	1.37	0.97	0.59	0.73

Results indicate that the normalized RMS error is smaller and the correlation coefficient is higher for articulatory features compared to raw movement data under the same inversion system, suggesting that the proposed palate-referenced features are better choices for representing the vocal tract configuration.

In Chapter 3, the variances of the original and palate-referenced features are compared. Figure 3.7 and table 3.3 shows that the vertical variance is significantly reduced in the palate-reference feature space, and figure 3.8 shows that the proposed features have significantly less overlap between the working spaces, strongly suggesting that the new features have better discriminatory representations than direct kinematic data. This directly influences the performance of HMM based acoustic-to-articulatory inversion due to increased separation between the observation distributions of different models, as shown by the decreased inversion error and increased correlation to actual

feature trajectories. From the inversion results the average decrease in normalized RMS error for the vertical dimension is 29% and the increase in correlation is 20%. These results strongly support the hypothesis that palate-referenced articulatory features are significantly more representative of vocal tract structure and acoustic spectral characteristics than direct sensor measures. Overall, the palate-referenced features have reduced variance and increased separation between vowels spaces and substantially lowered inversion error compared to direct sensor measures.

4.6 Summary

In this chapter a baseline HMM based inversion system has been built and evaluated. Acoustic and articulatory HMMs are trained independently, and articulatory parameters are recovered from the concatenated articulatory state sequences derived from forced alignment of acoustic model. A maximum likelihood parameter generation algorithm is used to produce trajectory output from a sequence of single Gaussian distribution. Additionally, two aspects of the system have been investigated for impact on inversion system performance: acoustic model complexity and dynamic window effect. By increasing the number of mixtures we improve the inversion performance, however, we need to monitor the performance across mixtures in order to avoid over fitting. Experimental results showed that for our data 7 mixtures gives the best performance in terms of average normalized RMS error and correlation. The other factor affecting the inversion performance is the selection of dynamic window coefficients. We investigated two commonly used windows, and the results show that the short-length 3-frame window based on theoretically optimal 1^{st} and 2^{nd} order regression coefficients gives the best

performance for recovering the slowly changing articulatory trajectory. In addition, the inversion performance between direct sensor movement and the palate referenced articulatory features proposed in chapter 3 have been compared. Results show that the palate referenced articulatory features have higher inversion accuracy, which supports our hypothesis that they better characterize the shape of the vocal tract.

The inversion model described in this chapter is a speaker dependent model requiring kinematic training data for each specific speaker. All experiments use a single male subject's data from EMA-MAE dataset. The next chapter will apply a model based speaker adaptation approach to extend this inversion system to work in a speaker independent domain.

5 Parallel reference speaker weighting for speaker independent inversion

5.1 Introduction

Most acoustic-to-articulatory inversion methods use parallel acoustic and articulatory training data from a single subject to learn the mapping between acoustic and articulatory spaces and then perform inversion on the acoustic data of the same subject. The mapping from the acoustic to the articulatory space varies across subjects due to physiological vocal tract differences, variability in speech production mechanisms, and differences in kinematic sensor placement across subjects. Therefore existing approaches for inversion are unlikely to work well if articulatory data from subjects are not available, as is realistically the case with many possible applications, such as Computer Aided Language Learning (CALL) or Computer Aided Pronunciation Training (CAPT) systems. An efficient speaker independent acoustic to articulatory inversion procedure needs to be developed which can estimate an unknown speaker's articulatory information from models trained using only from his or her acoustic realization.

There is significant evidence to suggest that multiple articulatory configurations can be associated with the same acoustic result (Atal et al., 1978; Lindblom et al., 1977; Qin & Carreira-Perpinan, 2007). It is nearly impossible to identify fine differences in articulatory configuration from the acoustic signal using existing methods. Within a single speaker creating an acoustic-articulator mapping is reasonable, but for multiple speakers it is a much more difficult problem. Because the relation between articulation

and acoustics is complex and non-linear, the problem cannot be solved with simple articulatory space and feature normalization. A method needs to be developed that will incorporate multiple acoustic-articulator mappings and create a new mapping that will be appropriate for a new speaker without reference kinematic data.

By using acoustic adaption techniques, the differences in acoustic patterns can be identified, and adapted acoustic and kinematic models in parallel can be created, to form a new inversion mapping that can estimate articulatory trajectory on new speakers. In this chapter a novel speaker independent inversion: parallel reference speaker weighting (PRSW) is developed and implemented based on the inversion system described in Chapter 4. Speaker dependent models for each subject enrolled in the experiments will be learned directly from the matched acoustic-articulatory data. Acoustically adapted models for each speaker will be created using the proposed PRSW method, using a target speaker's acoustic adaptation data without any kinematic data to determine PRSW weights and constructing a paired articulatory inversion model from the reference speakers. Each speaker will thus have measured articulator data as well as both speaker-dependent inversion model estimates and PRSW adapted kinematic-independent inversion model estimates. Direct evaluation of the acoustic-articulator inversion model will be done using correlation between actual and estimated articulatory features. The PRSW adaptation method will be discussed in 5.2, followed by experiments and results analysis in 5.3 and 5.4, respectively, with conclusions in 5.5.

5.2 Parallel Reference Speaker Weighting (PRSW)

As discussed in Chapter 2, Reference Speaker Weighting (RSW) is a rapid speaker adaptation approach that creates a new speaker model as a weighted combination of reference speakers, learning the appropriate weights from a small amount of adaptation data.

In PRSW, the speaker combination that generates the new speaker in acoustic space is assumed to be consistent with those in the articulatory space. The new speaker's articulatory realization can be recovered from the reference speakers' articulatory model by using acoustically derived weights. In the inversion stage, identical weights are used in the articulatory space. Let $\mathrm{A} = \{\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_\mathrm{K}\}$ be the set of reference speaker articulatory super vectors. Then the RSW estimate of the new speaker's articulatory supervector is

$$A_{unknown} \approx \sum_{k=1}^{K} w_k a_k = AW \tag{5.1}$$

W is the same weight derived from acoustic RSW in equation (2.12). The new speaker's articulatory movement can be estimated from the adapted model by using the maximum likelihood parameter generation algorithm described in section 4.3.3. Figure 5.1 illustrates this method for constructing an acoustic-articulator inversion model using the new PRSW approach.

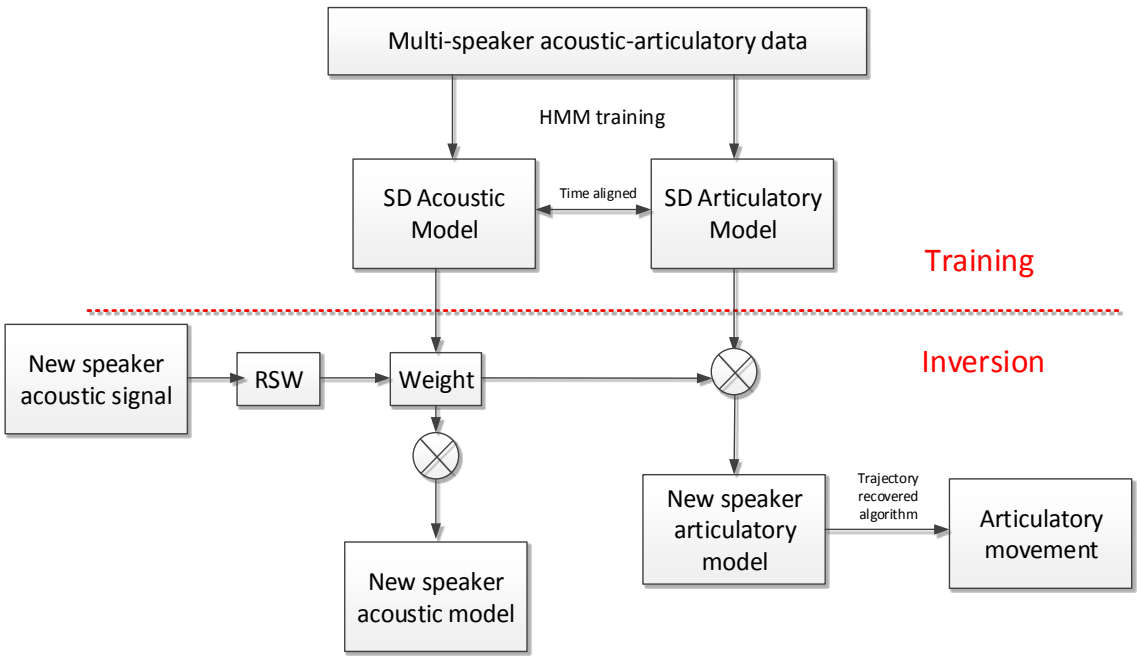

Figure 5.1 Diagram of Parallel Reference Speaker Weighting

This figure illustrates the method of constructing a speaker independent acoustic-to-articulatory inversion model. Using the multi-speaker articulatory and acoustic data, each reference speaker's parallel acoustic and articulatory HMMs are trained. Then RSW is used to adapt a new acoustic model for the unknown speaker. The weights derived from acoustic adaptation are combined in the same way in the articulatory space to generate the new speaker's articulatory model.

In order to evaluate the performance of the proposed PRSW, three different models have been implemented as in Figure 5.2.

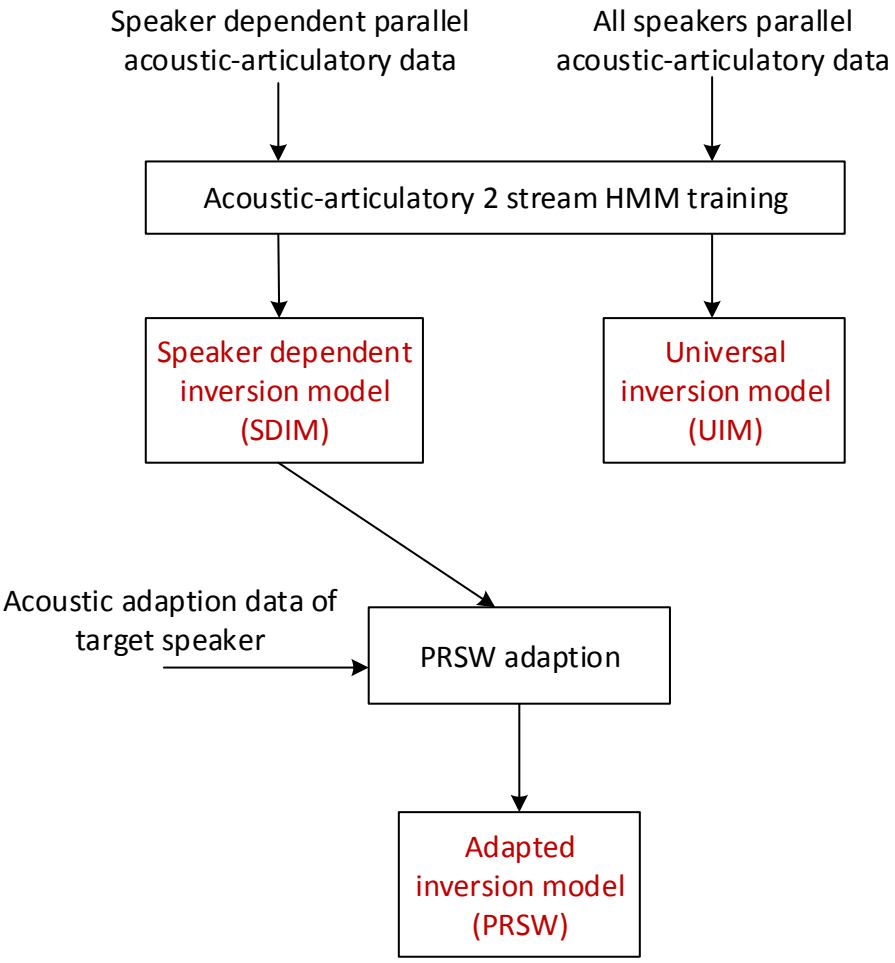

Figure 5.2 Implementation diagram of the three different models

Specifically, speaker dependent inversion models (SDIM) are trained on 45 minutes (including isolated words, sentences and paragraphs) of acoustic and parallel kinematic articulatory data for each speaker. A universal speaker independent inversion model (UIM) is trained on all speakers' data. The proposed PSRW method has also been implemented to get an adapted inversion model for each speaker only using acoustic data.

5.3 Experimental set up and evaluation

5.3.1 Experimental set up

20 Native American English speakers' data from the EMA-MAE dataset have been used in the following experiments. The baseline is the HMM based inversion system described in section 4.3 . A set of experiments has been performed to assess the PRSW model results:

1. A baseline adaption experiment to compare the inversion performance of SDIM, UIM and PRSW for each of the 20 speakers. The baseline experiment takes all available speakers' information into account to create a new speaker's inversion model**.** For each speaker, we excluded its own model from the 20 SDIM's pool and use the remaining 19 as reference speakers to estimate weights, then generate an adapted inversion model. The full 45 minutes of acoustic data for the target speaker is enrolled as adaptation data.
2. Reference speaker selection experiments to investigate the impact of the selection of reference speakers. Weight thresholding and global M-best pre-selection approaches will be compared and analyzed.
3. An experiment varying the amount of acoustic adaptation data to investigate data requirements of the different models in terms of inversion performance.

5.3.2 Evaluation

Normally, both average normalized RMS error and correlation are used to evaluate the performance for speaker dependent acoustic-to-articulatory inversion systems. In chapter 4, both of these are used to evaluate baseline inversion system. However, for a speaker independent framework, several studies (Ghosh & Narayanan, 2011; Hueber, Bailly, Badin, & Elisei, 2013) have shown that average normalized RMS error is not suitable for evaluating the cross-speaker acoustic-to-articulatory inversion due to differences in scaling and dynamic range caused by a lack of kinematic data. Without articulatory data for the test speaker the estimated articulatory outputs represent the correct movement patterns but not necessarily the new speaker's articulatory mean and variance, which are impacted by both physiological differences and sensor placement differences across subjects. Thus the correlation metric, which is a measure of overall similarity between the reference and the estimated trajectories, is a more appropriate evaluation criterion for quality of cross-speaker inversion results.

The correlation will be used to evaluate the inversion performance under different systems. Specifically, a set of experimental comparisons have been conducted to evaluate the proposed PRSW adaptation method:

1. Comparing the SDIM, UIM and PRSW inversion performance on the baseline experiment across 20 native English speakers. The hypothesis is that the adapted model should have better inversion output than the universal model and very close to the speaker dependent model.

2. Comparing the SDIM, UIM and PRSW inversion performance by applying different selection methods for reference speakers. The hypothesis is that the quality and articulatory consistency of the reference speaker set in will have impact on the PRSW performance, and thus that further improvement from using a reduced set of or relevant reference speakers.
3. Comparing the SDIM, UIM and PRSW inversion performance as a function of the amount of adaptation acoustic data. The hypothesis is that PRSW, being based on a rapid adaptation approach, is able to create accurate models using only a small amount of adaptation data.

5.4 Results and analysis

5.4.1 Baseline adaption result

Figure 5.3 shows the inversion performance for all 20 speakers in terms of correlation. From the correlation results we see that 13 out of 20 speakers support the initial hypothesis (SDIM > PRSW > UIM); however, 7 speakers have results that show a different pattern (SDIM > UIM > PRSW), with the PRSW method giving relatively poor results. If closely looking at the correlation, the inversion performance of the speaker dependent models varies widely across the 20 speakers (from highest 0.72 to lowest 0.52). The universal model has a relatively consistent inversion performance for every individual speaker (around 0.54).

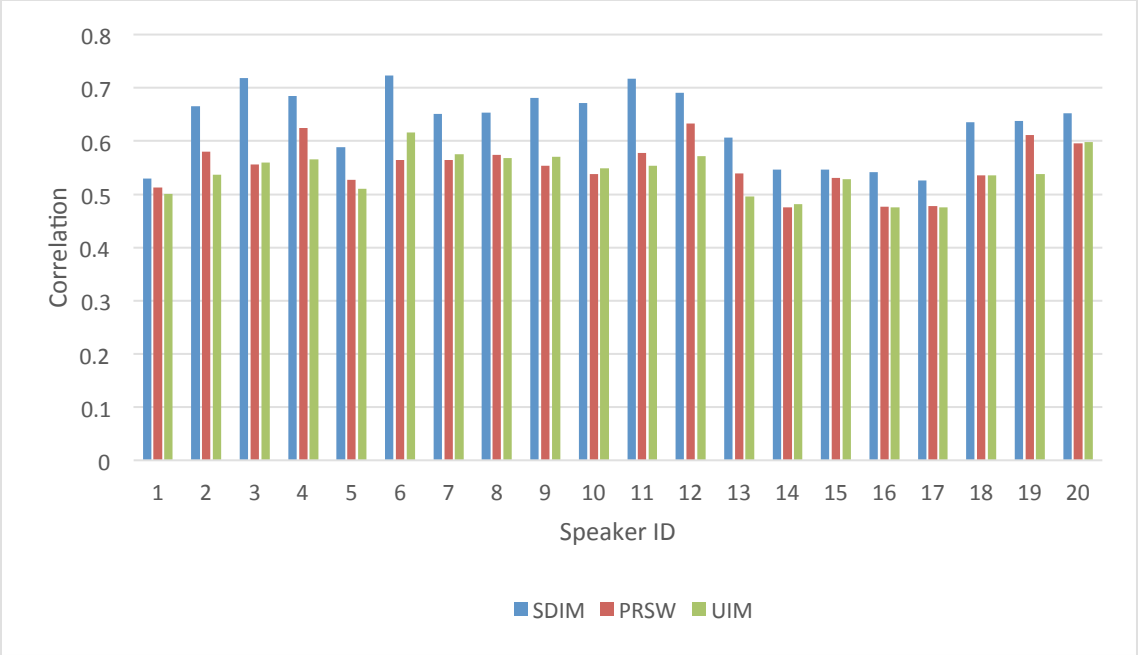

Figure 5.3 Baseline correlation results of the three different models

Figure 5.4 shows the average normalized RMS error for each speaker. The PRSW model always has the highest normalized RMS error.

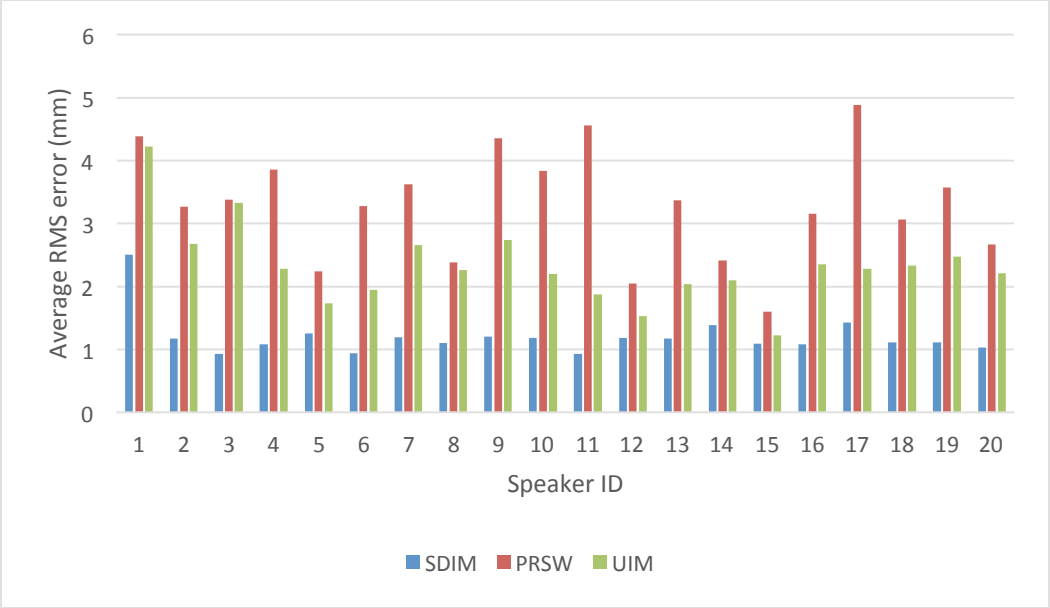

Figure 5.4 Baseline inversion results of three different models (normalized RMS error)

Looking in more detail at the RMS results, Figure 5.5 below illustrates why normalized RMS error is not a good measure for evaluating speaker independent inversion systems.

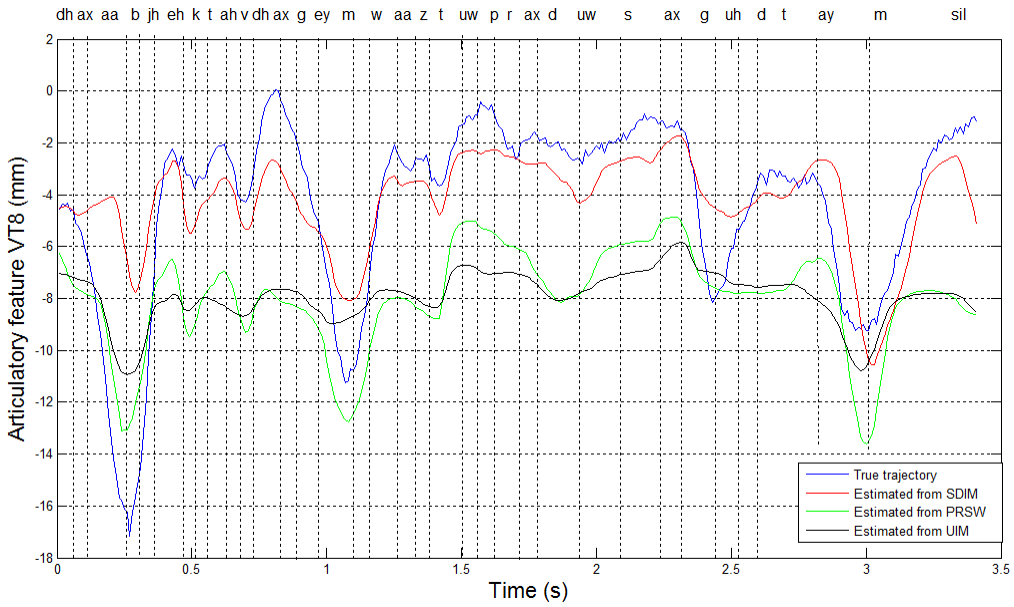

Figure 5.5 Recovered articulatory feature from the three different models

This figure shows the true trajectory (blue line) along with the inversion output from the SDIM (red line), the UIM (black line), and the PRSW model (green line) for the articulatory feature VT8. The average normalized RMS errors are 0.72, 1.25, and 1.32, respectively. The correlations are 0.78, 0.60, and 0.73, respectively. Although PRSW has the highest normalized RMS error, it is clear, both visually and from the correlation result, that it follows the shape of red and blue line much better than the UIM. The PRSW results show an offset of about 2.6 mm as observed in this figure. This is caused by physical variation between subjects, and there is no way to estimate or compensate for the offset without any articulatory information. Comparing figures 5.4 and 5.5 here supports the idea that average normalized RMS error is not suitable for evaluating the cross-speaker acoustic-to-articulatory inversion, as discussed previously in 5.3.2.

It should also be noted that correlation is a more meaningful measure with respect to most practical applications of acoustic-to-articulatory inversion. For speech recognition systems, articulatory synthesis systems, or pronunciation evaluation systems, the overall articulatory pattern is much more meaningful than exact sensor values. In fact, the speaker-independent approach shown here, using no kinematic data provides an implicit normalizing effect that acts to reduce speaker variance while still accurately tracking articulatory patterns, which would in many senses be expected to improve usefulness of the articulatory data to the target application.

5.4.2 Variation across speakers

From the baseline experiment results, there is a large variation in the original speaker dependent inversion performance across the 20 speakers. This variation can be further investigated by analyzing the articulatory feature model parameters for each speaker. The mapping from acoustic-to-articulatory space is through state alignment, so the more consistent the articulatory feature values are for identical phoneme sequences, the better the expected performance of the inversion system. The Gaussian variance in the articulatory HMM states are a good measure of this consistency.

The scatter plot in Figure 5.6 shows a linear relationship between the consistency of articulatory features and the inversion performance as measured by correlation. In this figure, each red dot represents an individual speaker. A higher variance indicates that the speaker has a less consistent articulatory pattern, which is correlated with the inversion model having less accurate estimates of articulatory feature patterns. Speakers with lower variance articulatory models have better performing inversion models.

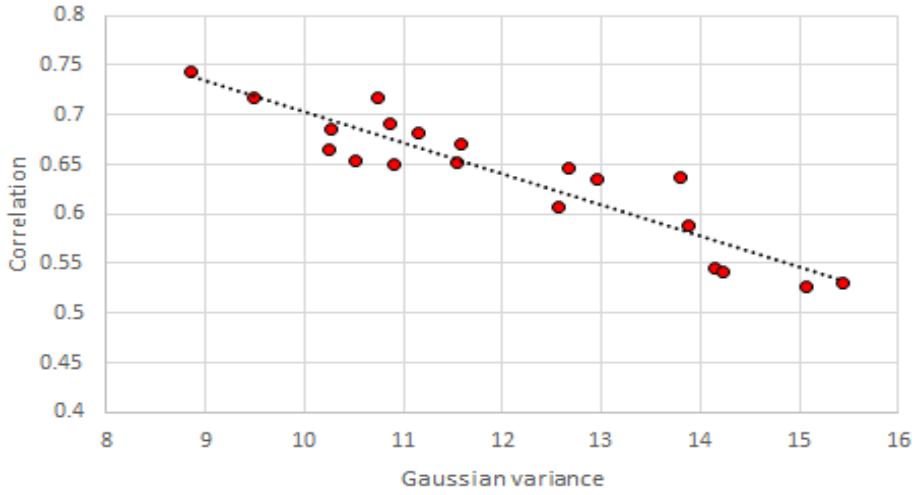

Figure 5.6 Scatter plot of articulatory model variance vs. correlation of speaker dependent models for all speakers

Cross-referencing the results from Figure 5.3 and Figure 5.6 and empirically reviewing the reference speaker weights reveals that the speakers with poor PRSW results are those whose primary reference speakers (highest weights in Equation 5.1) have high-variance speaker dependent models. This leads us to consider limiting the reference speaker set might be a way to improve the PRSW model.

In the next section, two different reference speaker selection strategies will be explored: one based on limiting the total number of reference speakers based on acoustic similarity (Weight thresholding) and the other based on globally limiting the reference speaker set based on speaker dependent inversion performance (M-best pre-selection).

5.4.3 Selection of reference speakers

Normally, the quality of an adapted acoustic model is dependent on the selection of reference speakers. The influence of selection approaches has been investigated in previous studies for acoustic models (Huang, Chen, & Chang, 2002; Kuhn et al., 2000) but not for articulatory models. In this section, two different reference selection strategies for the proposed acoustic-to-articulatory inversion system have been implemented and analyzed.

5.4.3.1 Weight thresholding

Figure 5.7 shows a diagram of the weight thresholding approach, based on acoustic model similarity. The RSW weights can be regarded as a similarity measurement, so that the best speakers in a nearest neighbor sense can be selected by setting a threshold α on sorted weights.

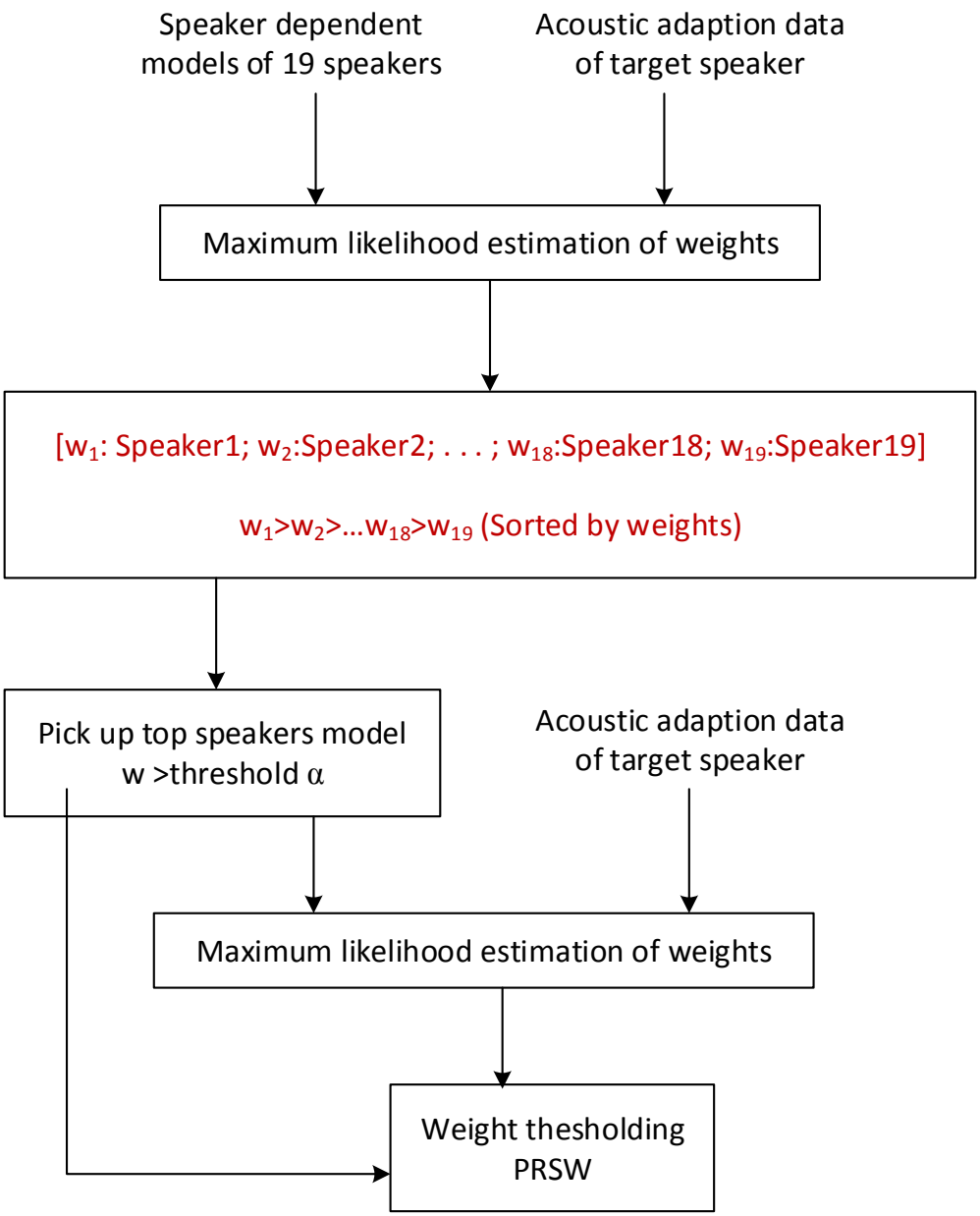

Figure 5.7 Weight thresholding PRSW

In order to investigate the effect of different thresholds, the threshold α is incremented in small steps (0.01), with maximum value of 0.09 to make sure that there is

at least one speaker in the reference speaker set. Figure 5.8 shows the plot of threshold as average performance across 20 speakers.

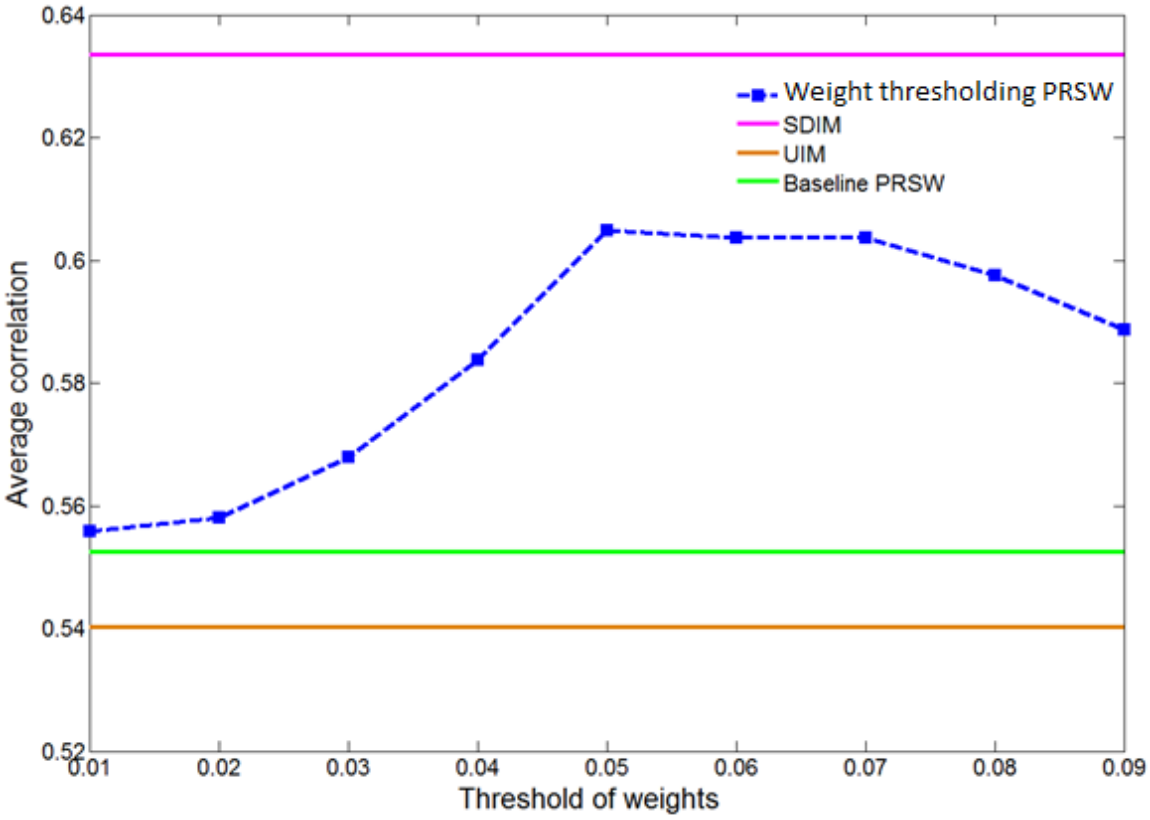

Figure 5.8 Plot of correlation as a function of threshold, for weight thresholding PRSW

This figure shows that the average performance curve is always higher than the baseline PRSW results. With the initial threshold of 0.01 the performance is close to that of the baseline PRSW. As the threshold increases, the performance continues to improve

until a threshold of 0.05, and then decreases slowly. The high performance suggests that reducing the number of speaker models being combined to create the new test speaker has a positive overall impact on articulatory consistency. Although in this case a threshold of 0.05 is the best, the optimal weight threshold would vary as a function of the original number of references speakers in different datasets. Once a specific reference speaker set is established, the optimal weight threshold can be determined on a set of development data and should give consistent results for new unseen speakers.

Table 5.1 shows the results using a threshold of 0.5 across all 20 speakers. The correlation performance shows significant improvement compared to the baseline PRSW system for each speaker.

Table 5.1 Weight thresholding PRSW results for all 20 speakers with the threshold ($\alpha = 0.05$)

Speaker ID	UIM	SDIM	Baseline PRSW	Thresholding	
				Correlation	M best (weight <0.05)
1	0.50	0.53	0.51	0.55	8
2	0.54	0.67	0.58	0.62	6
3	0.56	0.72	0.56	0.63	6
4	0.57	0.69	0.62	0.63	6
5	0.51	0.59	0.53	0.61	6
6	0.62	0.72	0.56	0.67	7
7	0.57	0.65	0.56	0.63	7
8	0.57	0.65	0.57	0.61	5
9	0.57	0.68	0.55	0.63	6
10	0.55	0.67	0.54	0.62	8
11	0.55	0.72	0.58	0.63	6
12	0.57	0.69	0.63	0.66	6
13	0.50	0.61	0.54	0.61	5
14	0.48	0.55	0.48	0.54	8
15	0.53	0.55	0.53	0.53	7
16	0.48	0.54	0.48	0.51	7
17	0.48	0.53	0.48	0.51	7
18	0.54	0.64	0.54	0.60	8
19	0.54	0.64	0.61	0.62	7
20	0.60	0.65	0.60	0.67	7
Average	0.54	0.63	0.55	0.60	6.65

Overall figure 5.8 and table 5.1 indicate that the adaptation model generated using the proposed weight-thresholding selection method achieves better inversion performance for unseen speakers compare to the baseline PRSW.

5.4.3.2 M-best global pre-selection

In this approach, the M speakers with the best speaker dependent inversion performance are selected globally as reference speakers, with the other speakers eliminated from consideration. Because of the observed large variance across speakers in terms of the quality of speaker dependent results, using inversion performance can be regarded as a measure of model consistency. The hypothesis is that the more consistent the reference speakers, the higher the upper limit on inversion results of the adapted model. Figure 5.9 details the M-best pre-selection approach based on speaker dependent inversion performance.

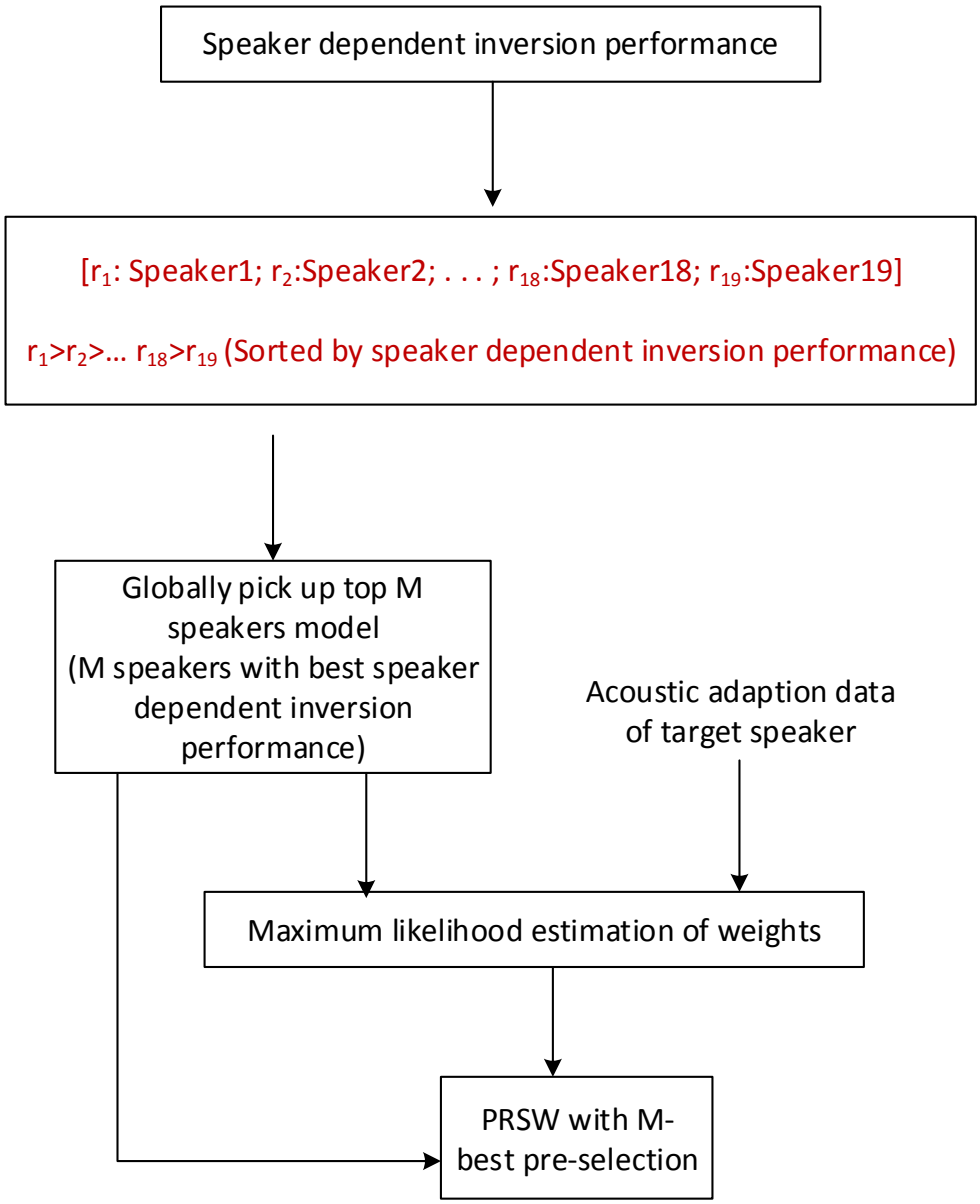

Figure 5.9 Diagram of PRSW with M-best pre-selection

In this global pre-selection method, the core reference speaker set is the same for each test speaker, including exactly the M-best reference speakers according to speaker dependent model correlation performance. When the test speaker is in the M best list, the

next best speaker is included instead, so that the reference set is maintained at M consistently across all 20 speakers. This means that the reference speaker sets are not fully identical, but always have at least 19 speakers in common. In this experiment, M is increased from 1 to 19. Figure 5.11 shows the plots of the average performance as a function of M across 20 speakers.

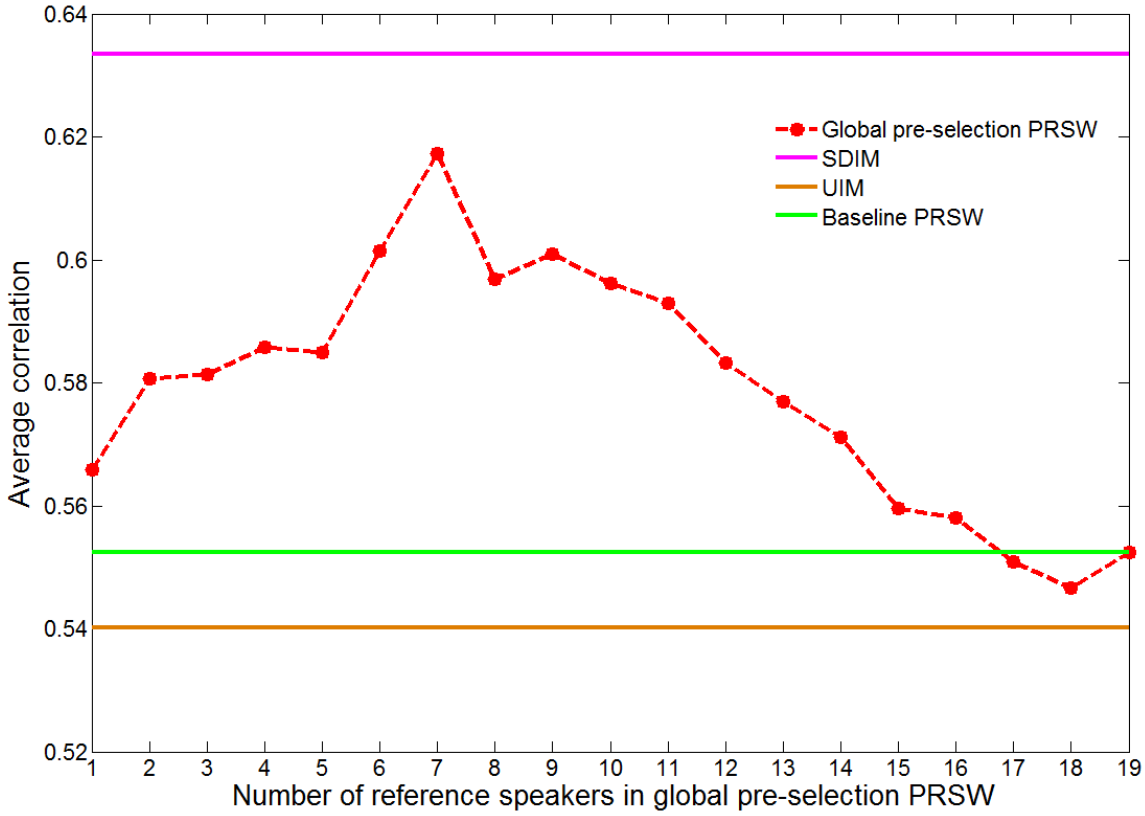

Figure 5.10 Plot of the inversion correlation results as a function of the number of reference speakers in M-best global pre-selection PRSW

The overall performance curve is above the baseline PRSW for M from 1 to 16. In the initial case M=1, a single reference speaker acts as a surrogate model for the target speaker. As the number of reference speakers increases, the average performance increases until reaching a peak at M=7, then decreases significantly. For this dataset, M = 7 results in only speakers having an SDIM correlation greater than 0.67 being selected as reference speaker in this case. As with the weight thresholding approach, the optimal parameter M is also a function of the original number of speakers, and more importantly of the quality of those speaker models as measured by the speaker dependent inversion performance.

Table 5.2 shows the results of the global M-best pre-selection PRSW approach for each individual speaker, with M=7. Results show that there is a large variation in the performance across 20 speakers, showing improvement over the baseline PRSW model in every case.

Table 5.2 Inversion correlation for each individual speaker using global M-best pre-selection PRSW with M=7

Speaker ID	UIM	SDIM	Baseline PRSW	M-best pre-selection PRSW M = 7
1	0.50	0.53	0.51	0.56
2	0.54	0.67	0.58	0.66
3	0.56	0.72	0.56	0.62
4	0.57	0.69	0.62	0.64
5	0.51	0.59	0.53	0.62
6	0.62	0.72	0.56	0.68
7	0.57	0.65	0.56	0.64
8	0.57	0.65	0.57	0.66
9	0.57	0.68	0.55	0.64
10	0.55	0.67	0.54	0.63
11	0.55	0.72	0.58	0.65
12	0.57	0.69	0.63	0.66
13	0.50	0.61	0.54	0.61
14	0.48	0.55	0.48	0.57
15	0.53	0.55	0.53	0.55
16	0.48	0.54	0.48	0.54
17	0.48	0.53	0.48	0.53
18	0.54	0.64	0.54	0.62
19	0.54	0.64	0.61	0.62
20	0.60	0.65	0.60	0.65
Average	0.54	0.63	0.55	0.62

The results show significant improvements compare to the baseline PRSW system, but with a large variation in terms of the amount of improvement. This difference might cause by the variation in both acoustic and articulatory patterns for each speakers. Having a globally reduced speaker set increases the likelihood that there will not be as many

good matches between the test speaker and the reference speaker set, in contrast to the weight thresholded approach where all speakers were initially included.

Table 5.3 shows the average performance for each of the proposed selection methods. Overall, both of these speaker selection approaches showed significant improvement compared to the baseline PRSW. Final results show that the M-best pre-selection PRSW gives the best inversion performance over this dataset.

Table 5.3 Comparison of inversion correlation performance

	Correlation
SDIM	0.63
M-best pre-selection PRSW (M=7)	**0.62**
Weight thresholding PRSW ($\alpha = 0.05$)	0.60
Baseline PRSW	0.55
UIM	0.54

By looking at the individual reference speakers selected in both acoustic and global selection method, it is interesting to find that there is a large overlap of the reference speakers' selection. The accuracy of the adapted model depends both on the similarity in the acoustic space and on the consistency of reference speakers articulatory patterns, but the latter is especially important. These two factors combined together affect the performance of adapted model. The results shown here strongly indicate that one of the biggest factors in high quality speaker independent kinematic-free acoustic-to-

articulatory inversion is a diverse set of reference speakers with consistent articulatory patterns.

5.4.4 Quantity of adaptation data

The PRSW experiments in the previous sections use the full set data from the target speaker to do adaptation, including 198 utterances representing approximately 28 minutes of speaking time. Normally RSW performs effectively when the amount of adaptation data is limited. One question is whether PRSW still has this property under our proposed inversion framework, and how much adaptation data is sufficient enough to obtain a good adapted articulatory model. In this section, the impact of amount of adaption data on the inversion performance is investigated. In the following experiments, the utterances set has been divided into 10 subsets. Table 5.4 shows the number of utterances in each subset.

Table 5.4 Number of utterances in adaptation subset

Adaptation Subset	1	2	3	4	5	6	7	8	9	10
Number of utterance	20	40	60	80	100	120	140	160	180	198

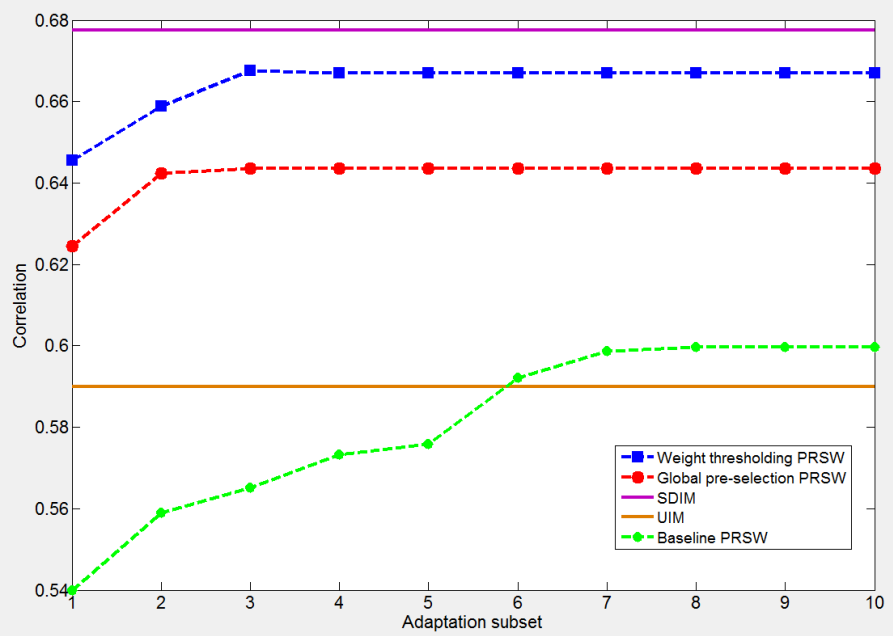

Figure 5.11 Inversion performance .vs. total quantity of adaptation data. (Each subset represents approximately 3 additional minutes of data)

Figure 5.11 shows the inversion performance versus the quantity of adaptation data for one speaker. The proposed PRSW method clearly shows a 'rapid' adaptation property compare to the full size of acoustic adaptation data. PRSW based on the reference speaker selection methods each converge at about 60 utterances while the baseline PRSW performance converges at 140 utterances. This can also be explained in relationship to the number of reference speakers used in each of the adaptation methods. In the baseline PRSW, 19 speakers are enrolled as reference speakers, thus more adaptation data is needed for the target speaker to estimate the ML weights. But the

reference selection methods decrease the amount of data needed for adaptation through compacting the size of reference speaker set.

This experiment has also been implemented for all 20 speakers individually. Results show the same rapid adaption property with slightly different converge points for each speaker, ranging from 20 to 80 utterances.

5.5 Summary

This chapter has presented a speaker independent acoustic-to-articulatory inversion system. A rapid speaker adaption approach, RSW, has been extended into the proposed PRSW framework in order to adapt the articulatory model from the acoustic space. The overall correlation between the true and estimated trajectories has been used to evaluate this system. In the initial baseline experiments, 13 out of 20 speakers' inversion results show that the adapted model using PRSW is better than the speaker independent model and very close to the speaker dependent model. Using the relationship between consistency of articulatory models and speaker dependent inversion performance, we investigate two new reference speaker pre-selection methods based on PRSW. Specifically, these methods include one based on thresholding the number of reference speakers based on acoustic model similarity, and another that is based on reducing the overall reference speaker set using speaker dependent inversion performance. Experimental results show that both of these selection methods work better than the baseline system, with significant improvements compared to the speaker independent model for each speaker. This indicates that the proposed PRSW is able to adapt a good articulatory model for the target speaker without any kinematic data as long as the

reference speaker set is carefully selected for acoustic and articulatory consistency. In addition, the impact of the amount of adaptation data on the inversion performance was investigated. The results show that the proposed PRSW method still preserves the rapid adaptation property in which the inversion performance converges with a small amount of adaptation data. Given a strong reference speaker set, the proposed PRSW adaptation is an effective approach for the speaker independent acoustic-to-articulatory inversion system even in the absence of kinematic training data.

6 Conclusions and future work

6.1 Contributions

Acoustic-to-articulator inversion, the estimation of articulatory trajectories from an acoustic signal, is an important problem with applications to a wide variety of speech processing technologies It is also a challenging problem due to the complexity of articulation patterns and significant inter-speaker differences. This is even more difficult when applied to speakers without kinematic training data.

The focus of this dissertation is solving the problem of acoustic-to-articulatory inversion when there is no kinematic data available. In order to achieve this goal, I have proposed and implemented a robust normalized articulatory space, a set of palate referenced articulatory features to model the vocal tract structure, and a novel speaker independent inversion system PRSW. To do this, existing model based speaker adaption methods used for speech recognition have been extended into the articulatory space. Specifically, a reference speaker weighting (RSW) approach has been utilized to identify differences in acoustic patterns and create adapted acoustic and articulatory models in parallel. This creates a new inversion mapping that can estimate articulatory trajectories on new speakers for whom there is limited acoustic adaptation data and no kinematic data. This study has achieved the following objectives:

1. The Marquette EMA-MAE corpus, a bilingual synchronized acoustic and kinematic data of 40 speakers, has been collected and used throughout the

dissertation. This dataset has been publically released to the research community for future research work in this area.

2. In chapter 3, a new articulatory space calibration method was introduced that includes head correction, bite plate calibration and palate surface estimation for the EMA-MAE corpus. The purpose of this calibration process is to transform the dataset into a meaningful anatomically referenced space, a normalized space that minimizes the difference across speakers.

3. Based on the new articulatory space and with the purpose of acoustic-to-articulatory inversion, a set of palate referenced articulatory features from EMA direct position measurements based on the vocal tract representation of Maeda's model is proposed. Direct working space comparison showed that the proposed articulatory features have smaller variance for the same vowel and more discrimination ability for different vowels within the same speaker. The proposed articulatory features have been evaluated using the baseline inversion system. The 29% average decrease in normalized RMS error and the 20% increase in correlation, compared to direct EMA sensor positions, strongly support the hypothesis that palate-reference articulatory features are significantly more representative of vocal tract structure and acoustic spectral characteristics.

4. The most important contribution is the method for speaker independent acoustic-to-articulatory inversion. The proposed Parallel Reference Speaker Weighting (PRSW) HMM-inversion system which can adapt to new speakers without any kinematic data has been implemented and tested. By adapting in

acoustic space, an adapted parallel articulatory model can be estimated to perform the inversion. Initial PRSW results on the EMA-MAE dataset, using a set of 19 reference speakers, produced an inversion accuracy close to that of the speaker dependent system for 13 out 20 speakers, By analyzing the inconsistency of inversion performance across speakers, we found that the accuracy of the adapted kinematic-independent models was related to the reference speaker basis set. This finding led me to investigate and implement two reference speaker selections approaches: one based on limiting the reference speakers individually based on acoustic similarity and the other based on globally limiting the total reference speaker set based on speaker dependent inversion performance. Experimental results show that both reference selection approaches obtained improvement compare to the baseline PRSW adapted model. The impact of the quantity of adaptation data on inversion performance was also investigated. Results show that the proposed PRSW is able to adapt a good articulatory model with relatively small amount of acoustic adaptation data. This suggests that adaptation for articulatory models requires somewhat more acoustic data compared to acoustic models due to the larger variation in articulatory space.

Overall, this study confirmed that articulatory patterns vary across speakers in consistent ways, ways that can be learned from associated reference speakers without needing kinematic data for each test speaker. The PRSW approach offers good speaker independent inversion performance without kinematic training data, but requires a set of reference speakers with consistent acoustic-articulatory patterns.

6.2 Future research

The proposed PRSW speaker independent acoustic-to-articulatory inversion has been implemented and evaluated across 20 native speakers. Based on the findings of this study there are several important directions for future research:

The EMA-MAE Marquette corpus includes another 20 Mandarin accented English speakers in addition to the 20 native speakers. It is important to investigate the difference in PRSW inversion performance between native and Mandarin accented speakers in the normalized working space with the proposed palate referenced articulatory. This comparison will provide direction on how to analyze non-native pronunciation patterns in articulatory space, and provide detailed corrective feedback to language learners.

Comparison of inversion performance within and across native and Mandarin accented speaker groups will also provide a good analyses across these two groups that will help characterize acoustic-articulatory relationships between mandarin and English speakers. Specifically, differences associated with vowels, consonant clusters, and contrastive stress variations should be analyzed and compared.

The PRSW solution may also be beneficial to the future development of CALL and CAPT systems, enabling them to provide specific corrective feedback mechanisms through direct assessment of articulatory movement by applying acoustic-to-articulatory inversion without the need for collecting kinematic data.

6.3 Conclusions

This dissertation introduces a novel speaker adaptation approach called Parallel Reference Speaker Weighting (PRSW), based on parallel acoustic and articulatory Hidden Markov Models. This approach uses a robust normalized articulatory space and palate referenced articulatory features combined with speaker-weighted adaptation to form an inversion mapping for new speakers to accurately estimate articulatory trajectories where there is no kinematic data. The proposed PRSW method is evaluated on the newly collected Marquette EMA-MAE corpus using 20 native English speakers. Cross-speaker inversion results show that given a good selection of reference speakers with consistent acoustic and articulatory patterns, the PRSW approach gives good speaker independent inversion performance, close to that of a speaker dependent system, without the need for kinematic training data.

References

Atal, B. S., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America, 63*(5), 1535-1555.

Badin, P., Bailly, G., Reveret, L., Baciu, M., Segebarth, C., & Savariaux, C. (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics, 30*, 533-553.

Bahl, L. R., & Jelinek, F. (1975). Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transaction on Information Theory, 21*, 404-411.

Beckman, M. E. J., & Jung, T. P. (1995). Variability in the production of quantal vowels revisited. *Journal of the Acoustic Society of America, 97*, 471-490.

Birkholz, P., Jackel, D., & Kroger, B. J. (2006). Construction and control of a three-dimensional vocal tract model. *International Conference on Acoustics Speech and Signal Processing,* 873-876.

Byrd, D., Browman, C. P., Goldstein, L., & Honorof, D. (1999). Magnetometer and x-ray microbeam comparison. *Proceedings of the 14th International Congress of Phonetic Sciences,* New York. 627-630.

Coker, C. H. (1976). A model for articulatory dynamics and control. *, 64*(4) 260-452.

Dang, J., & Honda, K. (2004). Construction and control of a physiological articulatory model. *Journal of the Acoustical Society of America, 115*(2), 853-870.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in contonously spoken sentences. *IEEE Transaction on Acoustics, Speech and Signal Processing, 28*(4), 357-166.

Dusan, S., & Deng, L. (2000). Acoustic-to-articulatory inversion using dynamical and phonological constraints. *In Proc. of the 5th Seminar on Speech Production: Models and Data,* Kloster Seeon, Germany. 237-240.

Erler, K., & Deng, L. (1993). Hidden markov model representation of quantized articulatory features of speech recognition. *Computer Speech and Language, 7*, 265-282.

Felps, D., & Osuna, R. G. (2010). *Normalization of articulatory data through procrustes transformations and analysis-by synthesis.* (Technical Report). Texas A&M University: Texas A&M University, Computer Science.

Frankel, J., & King, S. (2001). ASR-articulator speech recognition. *European Conference on Speech Communication and Technology,* Scandinavia.

TIMIT acoustic-phonetic continuous speech corpus. Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N. and Zue, V. (Directors). (1993). [Video/DVD] Linguistic Data Consortium.

Gauvain, J. L., & Lee, C. H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Acoustics, Speech and Signal Processing 2, 2*, 291-298.

Ghosh, P. K., & Narayanan, S. S. (2011). A subject-independent acoustic-to-articulatory inversion. *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference On,* 4624-4627.

Gracco, V. L., & Nye, P. W. (1993). Magnetometry in speech articulation research: Some misadventures on the road to enlightment. *Forschungber Institute Phonet., 31*, 91-104.

Hart, J. C., Francis, K. G., & Kauffman, H. L. (1994). Visualizing quaternion rotation. *ACM Transactions on Graphics, 13*(3), 256-276.

Hashi, M. Westbury, J. R., & Honda, K. (1998). Vowel posture normalization. *Journal of the Acoustical Society of America, 104*, 2426-2437.

Hazon, T. J. (2000). A comparison of novel techniques for rapid speaker adaptation. *Speech Communications, 31*, 15-33.

Hazon, T. J., & Glass, J. R. (1997). A comparison of novel techniques for instantaneous speaker adaptation. *Proceedings of the European Conference on Speech Communication and Technology,* 2047-2050.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America, 87*(4), 1738-1752.

Hiroya, S., & Honda, M. (2004). Estimation of articulatory movements from speech acoustics using an HMM-based speech production model. *IEEE Transactions on Speech Audio Process, 12*(2), 175-185.

Hiroya, S., & Mochida, T. (2005). Multi-speaker articulatory reconstruction based on an eigen-articulatory HMM. *IEEE International Conference on Acoustics, Speech and Signal Processing,* Philadelphia, PA, USA. 909-912.

Hofer, G., & Richmond, K. (2010). Comparison of HMM and TMDN methods for lip synchronisation. *Interspeech,* Makuhari, Japan. 454-457.

Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P., & Saltzman, E. (1996). Accurate recovery of articulator positions from acoustics: New conclusions based on human data. *The Journal of the Acoustical Society of America, 100*(3), 1819-1834.

Houde, R. A. (1967). *A study of tongue body motion during selected consonant sounds.* (PhD, University of Michigan).

Huang, C., Chen, T., & Chang, E. (2002). Speaker selection training for large vocabulary continuous speech recognition. *ICASSP,* Orlando, Florida, USA. 609-612.

Hueber, T., Bailly, G., Badin, P., & Elisei, F. (2013). Speaker adaptation of an acoustic-articulatory inversion model using cascaded gaussian mixture regressions. *Interspeech,* Lyon, France. 2753-2757.

IEEE subcommittee on subjective measurements IEEE recommended practices for speech quality measurements. (1969). *17*

Jelinek, F. (1969). A fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development, 13*, 675-685.

Jelinek, F. (1976). Continuous speech word recognition by statistical methods. *Proceedings of IEEE, , 64* 532-536.

Jelinek, F. (1999). *Statistical methods for speech recognition*. Cambridge, MA: MIT Press.

Jelinek, F., Bahl, L. R., & Mercer, R. L. (1975). Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory, 21*, 250-256.

Kaburagi, T., & Honda, M. (1994). An ultrasonic method for monitoring tongue shape and the position of a fixed-point on the tongue surface. *Journal of the Acoustical Society of America, 95*(4), 2268-2270.

Kaburagi, T., & Honda, M. (1998). Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory–acoustic database. *Proc. ICSLP,* Sydney, Australia. 433-436.

King, S., & Wrench, A. A. (1999). Dynamical system modeling of articulator movement. *International Conference on Phonetic Sciences,* San Francisco, USA.

Kirchhoff, K. (1999). *Robust speech recognition using articulatory information.* (PhD, University of Bielefeld).

Krista, R. (2011). *The effect of palate morphology on consonant articulation in healthy speakers.* (Master, Department of Speech-Language Pathology).

Kubala, F., Schwartz, R., & Barry, C. (1989). Speaker adaptation using multiple reference speakers. *DARPA Speech and Language Workshop,* San Mateo, CA.

Kuhn, R. (1998). Eigenvoices for speaker adaptation. *International Conference on Spoken Language Processing,* Syndey, Australia. 1771-1774.

Kuhn, R., Junqua, C., Nguyen, P., & Niedzielski, N. (2000). Rapid speaker adaptation in eigen voice space. *IEEE Transactions on Speech Audio Proceedings, 8*, 695-707.

Laprie, Y. (1998). A variational approach for estimation vocal tract shapes from the speech signal. *International Conference on Acoustic, Speech and Signal Processing,* Seattle, USA.

Lawrence, H., & Schafer, R. W. (1978). *Digital processing of speech signals* Prentice-Hall.

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language,* , 171-185.

Leung, K. Y., & Siu, M. (2004). Speech recognition using combined acoustic and articulatory information with retraining of acoustic model parameters. *International Conference on Spoken Language Processing,* Jeju Island, Korea.

Lindblom, B., Lubker, J., & Gay, T. (1977). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predicitive simulation. *The Journal of the Acoustical Society of America, 62*(S1), 1115-1123.

Ling, Z., Richmond, K., Yamagishi, J., & Wang, R. (2009). Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing, 17*(6), 1171-1185.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. Boston: Kluwer Academic Publishers.

Masaki, S., Tiede, M. K., Honda, K., Shimada, Y., Fujimoto, I., Nakamura, Y., & Ninomiya, N. (1999). MRI-based speech production study using a synchronized sampling method. *The Journal of the Acoustical Society of America, 20*(5), 375-379.

McGowan, R., & Cushing, S. (1999). Vocal tract normalization for midsagittal articulatory revovery with analysis-by-synthesis. *Journal of the Acoustical Society of America, 106*(2), 1090-1105.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America, 53*(4), 1070-1082.

Metze, F., & Waibel, A. (2002). A flexible stream architecture for ASR using articulatory features. *The International Conference on Spoken Language Processing,* Denver, USA.

Mitra, V., Nam, H., Espy-Wilson, Y., Saltzman, E., & Goldstein, L. (2010). Retrieving tract variables from acoustics: A comparison of different machine learning strategies. *IEEE Journal of Selected Topics in Signal Processing, 4*(6), 1027-1045.

Munhall, K. G., Vatikiotis-Bateson, E., & Tohkura, Y. (1998). X-ray film database for speech research. *Journal of the Acoustical Society of America,* , 1222-1224.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., & Byrd, D. (2004). An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America, 115*(4), 1771-1776.

Perkell, J. S., & Cohen, M. H. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America, 92*, 3078-3086.

Qin, C., & Carreira-Perpinan, M. A. (2007). An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. *Interspeech,* Belguim. 74.

Rabiner, L. R. (1989). Tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE. 77* 257-286.

Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Richmond, K. (2002). *Estimating articulatory parameters from the acoustic speech signal.* (PhD, The Centre for speech technology research, Edinburgh University).

Richmond, K., Hoole, P., & King, S. (2011). Announcing the electo-magnetic articulatography (day 1) subset of the mngu0 articulatory corpus. *Interspeech,* Florence, Italy. 1505-1508.

Rogers, C. L. (1997). *Segmental intelligibility assessment for chinese-accented english.* (PhD, University of Indiana).

Scobbie, J. M., Turk, A., Geng, C., King, S., Lickley, R. J., & Richmond, K. (2013). The ediburgh speech production facility doubletalk corpus. *Interspeech,* Lyon, France. 764-766.

Stone, M. L., Sonies, B. C., Shawker, T. H., Weiss, G., & Nadel, L. (1983). Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. *Journal of Phonetics, 11*(3), 207-218.

Story, B. (2005). Synergistic modes of vocal tract articulation for american english vowels. *Journal of the Acoustical Society of America, 118*, 3834-3859.

Sun, J., & Deng, L. (2002). An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition. *Journal of the Acoustical Society of America,* 1086-1111.

Tang, M., Seneff, S., & Zue, V. (2003). Modeling linguistic features in speech recognition. *European Conference on Speech Communication and Technology,* Geneva.

Toda, T., Black, A., & Tokuda, K. (2004). Acoustic-articulatory inversion mapping with gaussian mixture model. *International Conference on Spoken Language Processing,* Jeju Island, Korea. 1129-1132.

Tokuda, K., Yoshimura, T., Masuko, T., & Kobayashi, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. *ICASSP,* Istanbul. *, 3* 1315-1318.

Wei, J. (2008). Vocal tract normalization in articulatory space using thin-plate spline method. *Journal of Acoustical Society of America, 123*(5), 3885.

Westbury, J. (1991). The significance and measurement of head position during speech production experiments using the x-ray microbeam system. *Journal of Acoustical Society of America, 89*(4), 1782-1797.

Westbury, J. (1994a). In University of Wisconsin Press (Ed.), *X-ray microbeam speech production database user's handbook* (1st ed.). Madison: University of Wisconsin Press.

Westbury, J. (1994b). *X-ray microbeam speech production database user's handbook version 1.0*

Wrench, A. A. (1993). EUR-ACCOR corpus. Retrieved from http://www.cstr.ed.ac.uk/research/projects/artic/accor.html

Wrench, A. A., & William, J. (2000). A multichannel articulatory database and its application for automatic speech recognition. *5th Seminar on Speech Production: Models and Data,* Bavaria. 305-308.

Yunusova, Y., Baljko, M., Pintilie, G., Rudy, K., Faloutsos, P., & Daskalogiannakis J. (2012). Acquisition of the 3D surface of the palate by in-vivo digitization with wave. *Speech Communication, 54*(8), 923-931.

Zhang, L., & Renals, S. (2008). Acoustic-articulatory modelling with the trajectory HMM. *IEEE Signal Processing Letters, 15* 245-248.

Zue, V., Seneff, S., & Glass, J. R. (1990). Speech database development at MIT. TIMIT and beyond. *Speech Communication, 9*(4), 351-356.